STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:753–773
Published online 3 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.1947

The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering

Marco Bonetti[1,2,*,†] and Marcello Pagano[1]

[1]*Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.*
[2]*Dana-Farber Cancer Institute, Boston, MA 02115, U.S.A.*

SUMMARY

The topic of this paper is the distribution of the distance between two points distributed independently in space. We illustrate the use of this interpoint distance distribution to describe the characteristics of a set of points within some fixed region. The properties of its sample version, and thus the inference about this function, are discussed both in the discrete and in the continuous setting. We illustrate its use in the detection of spatial clustering by application to a well-known leukaemia data set, and report on the results of a simulation experiment designed to study the power characteristics of the methods within that study region and in an artificial homogenous setting. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: distance-based methods; Monte Carlo sampling; U-statistics; disease clusters

1. INTRODUCTION

Consider the distance between two points. If one of the points is fixed and the other random, then we have a non-negative random variable and a large scientific literature associated with its study. On the other hand, if both points are random, then the general study of such a random distance occupies only a rather small part of the statistical literature, and only in the simpler cases can its distribution be derived analytically (see References [1–4]). To draw inference about such a distribution, one may take a random sample of n points, which result in the larger (for $n>3$) number of $\binom{n}{2}$ dependent distances.

Except for very simple cases, it is very difficult to analytically express the dependencies among these distances. But yet it is informative, and thus desirable, as we show below, to study such distributions. Their natural estimator, the empirical frequency distribution

*Correspondence to: Marco Bonetti, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, U.S.A.
†E-mail: bonetti@jimmy.harvard.edu

Contract/grant sponsor: National Institutes of Health; contract/grant numbers: AI28076 and LM07677-01

Received May 2003
Accepted May 2004

function ('ecdf') of the $\binom{n}{2}$ dependent distances, can form the basis for inference. Because of the dependencies, the study of this estimator does not follow the usual paradigm of an empirical cumulative distribution function based on independent identically distributed (i.i.d.) observations, and thus it is not as straightforward to obtain its sampling properties.

There is a round about way of arriving at this estimator that follows more familiar lines. Suppose n is even. We can easily obtain $n/2$ independent distances, and construct their empirical cdf. There are $n!/(2^{n/2}(n/2)!)$ ways of choosing the $n/2$ independent distances. To gain efficiency we can take a resampling approach and average all possible empirical cdfs based on $n/2$ independent distances. It is not difficult to show that with this approach one recovers exactly the frequency distribution of all the dependent distances, the ecdf. A parallel may be drawn with the calculation of the sample variance S_n^2 of n (even) numbers $X_1,\ldots,X_n$. Given $S_n^2=(n-1)^{-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$, with $\bar{X}_n$ the sample mean, it is well known that $S_n^2=(n(n-1))^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(X_i-X_j)^2$, an average of dependent quantities. Considering a random permutation θ of the indices $i=1,\ldots,n$, one can then define the estimator $S_\theta^2=n^{-1}\sum_{i=1}^{n/2}(X_{2i-1}-X_{2i})^2$, an unbiased but inefficient estimator based on independent summands. Then averaging these S_θ^2 over all $n!/(2^{n/2}(n/2)!)$ distinct ways of creating independent pairs, yields exactly S_n^2.

The ecdf converges to the distribution of the interpoint distance between two randomly selected observations [3], so that for finite, but large n, one may compare the ecdf of the $\binom{n}{2}$ distances to its population counterpart to evaluate the agreement between the sample and a hypothesized population distribution. The genesis of the idea to use the interpoint distance distribution is evident in the work of Bartlett [2], who studies points uniformly distributed within a unit circle and a unit square. This approach is applicable to the situation in which the points are generated according to an absolutely continuous distribution over a region, as well as to the situation in which the points are constrained to belong to one of a fixed, and possibly finite, set of possibilities.

In what follows we will see that the ecdf of all pairwise distances evaluated at a finite number of values along the distance axis has an asymptotic multivariate normal distribution. More generally, we also provide a new proof of the result that the centred empirical frequency distribution of the pairwise distances converges to a Gaussian process. One can then evaluate the difference between the empirical frequency distribution and its population counterpart in a variety of ways. For example, if the ecdf is computed over a finite grid, then a statistic resembling a Mahalanobis distance can be used to construct a chi-squared-like test statistic. In Section 2, we discuss the interpoint distance distribution both in the continuous case and in the discrete case. The initial motivation for our work was the problem of the detection of disease clustering over a population non-uniformly distributed over a region, and in Section 3, we show an application of our methods to that particular setting with an illustration based on a well-known data set. In Section 4, we describe a simulation study of the power of the proposed methods in comparison to some other existing clustering statistics. This motivation for our work influences the assumptions we make of our models. In general, we view the sampling *region* as given and fixed, and not as a sampled part of a larger whole. As a consequence, for inference we eschew such restrictive assumptions as stationarity of underlying point processes and prefer to turn to exact resampling methods.

2. THE INTERPOINT DISTANCE DISTRIBUTION

2.1. The continuous case

Consider first a point process where the observations can appear anywhere inside some bounded region. Let the point distribution over the region be absolutely continuous, so that for two i.i.d. points X_1 and X_2 in the region, $\Pr(X_1 = X_2) = 0$.

For any point distribution P in a region S, on which is defined a non-negative distance (or dissimiliarity) function d, the cdf $F(\cdot)$ of the interpoint distance D between two independent points is $F(d) = \mathscr{E}1(d(X_1, X_2) \leqslant d)$, where $1(\cdot)$ is the indicator function and $\mathscr{E}$ denotes expectation with respect to the $P \times P$ distribution. For example, on the plane, Bartlett [2] reports the distribution of the interpoint distances for randomly distributed points on the unit square and on the unit circle (results originally due to Borel [1]), and he suggests computing a chi-square test to measure the deviation between the observed and the expected frequencies over a grid. He also recognizes that distributional problems arise because the observed distances do not constitute a sample of independent observations.

If one views the sampling region as itself a sample of some bigger space, then to extrapolate the results beyond the region we require some property of the process to make this generalization reasonable. One such property is that of stationarity. A point process defined on a topological space S is said to be *stationary* if its distribution is invariant under a topological group G acting continuously on S (a typical example being the group G of rigid motions acting on the plane). The definition, and use, of the interpoint distribution function $F(d)$ given above does not require that the point process be stationary, but if it is, a number of theoretical results are available. In the setting of stationary isotropic processes, Ripley [5] defines the K-function

$$K(t) = \lambda^{-1} E[\text{number of further events within distance } t \text{ of an arbitrary event}]$$

where λ is the *intensity* of the process, or the (assumed constant) expected number of events per unit of area. Ripley points out that the K-function shares some of the properties of the interpoint distribution function, even though it is not a distribution function; indeed, $K(t) \to \infty$ as $t \to \infty$. He proposes an estimator of $K(t)$ that in the case of the unit square is unbiased for $t < 1/\sqrt{2}$; i.e. half of the maximum distance observable in the unit square, and has variance that increases rapidly as t increases. Also, if we define $Y(t)$ to be the number of interpoint distances within a region S which are within t of each other (a non-normalized version of the ecdf), then Silverman and Brown [6, 7] prove the weak convergence of $Y(t)$ on $[0, t_0]$ when t_0 is *small* relative to the maximal distance in S. Within that small interval, $Y(t)$ converges to a heterogeneous Poisson process (see also [8, p. 44]).

Extending the usual definition of an empirical distribution function for random samples, we define the ecdf of the interpoint distances associated with a random sample $X_1, \ldots, X_n$ as

$$F_n(d) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 1(d(X_i, X_j) \leqslant d)$$

For fixed d, $F_n(d)$ is an example of a V-statistic (see for example Reference [9, p. 172]). In the appendix it follows that the scaled distribution of $F_n(d)$ computed at a finite set of values, $d_1, \ldots, d_m$, converges to a multivariate normal distribution as $n \to \infty$ (see also Reference [3] for an alternate proof).

Silverman [3] further showed that the quantity $\sqrt{n}(F_n(d)-F(d))$, considered as a stochastic process indexed by d, converges weakly to a Gaussian process. His proof can be shortened considerably by making use of recent results from the theory of U-processes (see Reference [10]), as shown in the appendix. The ecdf of the set of dependent interpoint distances among n points in the plane is thus a well-defined and behaved summary of a configuration of points. One characteristic of such a descriptor is its rotational invariance, a property that it shares with all distance-based statistics.

2.2. The discrete case

Consider now a region within which points (individuals) can arise at any of the fixed locations $l_1,\ldots,l_k$ with probabilities $p_1,\ldots,p_k$ (with $\sum_{j=1}^{k} p_j = 1$). Let the random variable D again represent the distance between two individuals chosen at random from this population. Let d_{ij} be the distance between locations l_i and l_j. The random variable D thus takes on the value d_{ij} with probability $p_i p_j$. The distribution function of this non-negative random variable is

$$F(d) = F(d;\, p) = \sum_{i=1}^{k}\sum_{j=1}^{k} p_i p_j 1(d_{ij} \leqslant d) \tag{1}$$

Consider a random sample $n_1,\ldots,n_k$ of individuals over this region, and let $n = \sum_{i=1}^{k} n_i$. Consider all the $\binom{n}{2}$ distances between the individuals in the sample, and compute the function $F_n(d) = F(d;\hat{p})$, where $\hat{p} = (\hat{p}_1,\ldots,\hat{p}_k)$ and for $i = 1,\ldots,k$, $\hat{p}_i = n_i/n$. Note how these definitions of $F(d;p)$ and $F(d;\hat{p})$ are the discrete analogues of $F(d)$ and $F_n(d)$ given in Section 2.1 for the general continuous case.

Since we are interested in the distribution of the distances between individuals, and we do not wish to make assumptions or inference about the value of the sample size, n, we condition on it. We can then use the distribution of the distances obtained by choosing samples of size n at locations l_i with probabilities p_i, $i = 1,\ldots,k$ (see Reference [11]) as the null distribution. Then the null hypothesis of random sampling from the population distribution is the hypothesis that the n_i are a multinomial sample with probabilities $p = (p_1,\ldots,p_k)$. Since the $\widehat{p}_i$ are strongly consistent estimators of the p_i (as $n \to \infty$), for any fixed and real d, $F(d;\hat{p})$ is a strongly consistent estimator of $F(d;p)$. A measure of the difference between $F(d;\hat{p})$ and $F(d;p)$ can thus be used as a gauge of the null hypothesis of spatial randomness.

Note that in this discrete setting (as opposed to the continuous case) one can expect the underlying population distribution to be known at least approximately. Here, also, for a fixed value d the empirical cdf $F(d;\hat{p})$ has $\sqrt{n}$-convergence to $\mathscr{E}(d(X_1,X_2) \leqslant d)$. Moreover, the convergence to a multivariate normal distribution holds when one computes the cdf at the finite set of values $d_1, d_2, \ldots, d_m$.

2.3. Test statistics

A large number of standard test statistics can be used to evaluate the distance between $\hat{F}_n(\cdot)$ and $F(\cdot)$, but the lack of independence between observed distances between individuals precludes the use of standard statistics without using appropriate modifications.

Just as one does for a histogram, one can define an increasing collection of values $\mathbf{d} = \{d_1,\ldots,d_m\}$ over the range of D and define the two vectors $F_n(\mathbf{d}) = \{F_n(d_1),\ldots,F_n(d_m)\}$ and $F(\mathbf{d}) = \{F(d_1),\ldots,F(d_m)\}$.

The asymptotic normality noted in the previous sections suggests the following statistic to measure the distance between $F_n(\mathbf{d})$ and $F(\mathbf{d})$:

$$\tilde{M}(F_n(\mathbf{d}), F(\mathbf{d})) = (F_n(\mathbf{d}) - F(\mathbf{d}))' \Sigma^- (F_n(\mathbf{d}) - F(\mathbf{d})) \tag{2}$$

a Mahalanobis-like statistic, where Σ^- is a generalized inverse (see Reference [12]) of the covariance matrix of the vector $F_n(\mathbf{d})$. For definiteness we use the Moore–Penrose generalized inverse. One can, in theory, compute the exact distribution of $\tilde{M}$, but if n is of any reasonable size, the calculation is not feasible. As an alternative, one could appeal to the asymptotic results in the appendix, but empirical experience suggests that the convergence of the distribution of $\tilde{M}$ to its asymptotic value is quite slow. This is especially so in discrete situations where there are many locations, since then typically a number of the probabilities p_i involved are small. Because of this, we do not use $\tilde{M}$, but rather propose using an estimator of $\tilde{M}$. Consider M defined as $\tilde{M}$, but with the estimated covariance matrix

$$M(F_n(\mathbf{d}), F(\mathbf{d})) = (F_n(\mathbf{d}) - F(\mathbf{d}))' S^- (F_n(\mathbf{d}) - F(\mathbf{d})) \tag{3}$$

where S is the sample covariance estimator obtained after taking repeated samples, with replacement, of size n. This is the statistic we propose to use, with the generalized inverse matrix S^- chosen to be the Moore–Penrose generalized inverse of S. In practice we have sampled repeatedly 1000 times with success.

When comparing the sample and theoretical distributions, it is sometimes more instructive to see the scaled first difference function $f_n(d)$

$$f_n(d) = \tfrac{1}{\varepsilon}[F_n(d + \varepsilon/2) - F_n(d - \varepsilon/2)]$$

One typically defines (and plots) a vector $f_n(\mathbf{d}) = (f_n(d_1), \ldots, f_n(d_m))$ of values computed at values $d_1, \ldots, d_m$ taken here to be such that $d_j - d_{j-1} = \varepsilon$ for $j = 1, \ldots, m$ and m some positive integer. We set $d_1 = \varepsilon/2$, and define $f_n(d_1) = F_n(\varepsilon)/\varepsilon$ so that it includes the origin. The population equivalent of $f_n(\mathbf{d})$ is the vector $f(\mathbf{d}) = (f(d_1), \ldots, f(d_m))$ computed at the same values $d_1, \ldots, d_m$, but replacing $F_n(\cdot)$ by $F(\cdot)$. Because of its linear relationship with $F_n(\cdot)$, the first difference function $f_n(\cdot)$ has $\sqrt{n}$-convergence to the expected value $\mathscr{E}(1(d - \varepsilon/2 < d(X_1, X_2) \leqslant d + \varepsilon/2)$, and for a fixed d, $n^{1/2} f_n(d)$ has an asymptotically normal distribution. The joint asymptotic distribution for multiple values of d also follows immediately.

Above we have defined the statistic M and $\tilde{M}$ in terms of $F(\cdot)$ and $F_n(\cdot)$, but note that we could equally well define them in terms of $f_n(\cdot)$ and $f(\cdot)$ computed at the same values $\mathbf{d} = (d_1, \ldots, d_m)$. The two forms with, of course, appropriate definitional changes in the covariance matrix, yield identical results. Statistics other than M can be defined by choosing a different distance measure between $f_n(\cdot)$ and $f(\cdot)$. Below we explore the following possibilities:

$$M_1(f_n, f) = \int_0^\infty (f_n(x) - f(x))^2 \, \mathrm{d}\mu(x)$$

$$M_{\chi^2}(f_n, f) = \int_0^\infty \frac{(f_n(x) - f(x))^2}{f(x)} \, \mathrm{d}\mu(x)$$

$$M_{\mathrm{KL}}(f_n, f) = \int_0^\infty \log\left(\frac{f_n(x)}{f(x)}\right) f(x) \, \mathrm{d}\mu(x)$$

M_1 is the L_2 norm of the difference between f_n and f; M_{χ^2} is a χ^2-type distance; and M_{KL} is the well-known Kullback–Leibler semi-metric. One method of approximately evaluating these integrals is with respect to the discrete measure μ that puts equal mass at equispaced points at which f_n and f are evaluated, so that they become sums. Derivation of the asymptotic distributions of these statistics is difficult, but we can again rely on the Monte Carlo sampling approach to construct tests of hypotheses.

3. AN APPLICATION TO CLUSTER DETECTION

3.1. Disease clustering

The search for clusters in the spatial distribution of a set of points is an important problem with a long history in statistics. One notable application of methodologies developed in this context is the search for disease clustering, especially in response to alarms raised by the public. (See for example Reference [13] and references therein).

In their search for such clusters, the Centres for Disease Control and Prevention in Atlanta have issued cluster detection guidelines that contain the rather pessimistic statement that 'in many reports of cluster investigations, a geographic or temporal excess in the number of cases cannot be demonstrated' [14]. This guarded view may be the result of the rather poor success rate in cluster detections (out of 108 suspected cancer clusters investigated over 22 years, no clear cause was found for any of them; see Reference [15]), although this could mean either that false alarms are raised too easily (alarms that under further study are readily dismissed), or that existing methods are not sufficiently powerful for detecting clusters. Some of the existing clustering methods are reviewed in Reference [16], where one can also find a description of many model-based approaches aimed at assessing dependence in spatial point processes.

We can use the methods described here to test whether the observed interpoint distance distribution among the individuals with a certain disease is consistent with the hypothesis of no disease-induced clustering. It should be noted that these methods are designed to detect any disturbance from such distribution, and not just a single cluster. This can be quite important, since in most cases one does not know the number, shape and location of the clusters that may exist. In this section we discuss the discrete case, because the application that follows is discrete, as it is in many cluster investigations since the data typically is only available in aggregate form; either because of the data collection method or because of concerns for confidentiality.

For a given set of fixed centres, the lack of deviations from the population distribution is equivalent to choosing as cases of the disease of interest individuals, at random, from the centres, with probabilities given by the appropriate population proportions: $p_i,\ i=1,\ldots,k$. So when we consider a group of individuals with a particular ailment (leukaemia, say), and ask whether they are geographically distributed as in the population—the null hypothesis of no clustering—then we immediately think of the goodness-of-fit problem, and the associated classical chi-squared test for the multinomial distribution. This test is a general one and is not targeted at the clustering problem at hand. Indeed, we can think of the chi-squared goodness-of-fit test as a quadratic form $(\hat{p}-p)'\Sigma^{-}(\hat{p}-p)$ involving the difference between the observed $(\hat{p})$ and expected (p) proportions, where the weighting matrix Σ^{-} is the generalized inverse

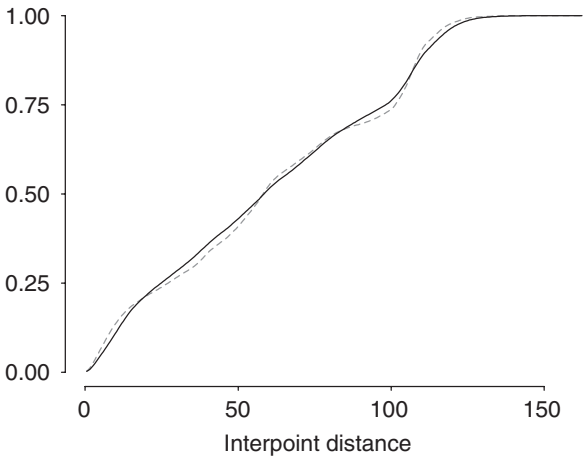

Figure 1. The solid line is the cdf of the distances between individuals reported in the 1980 census and the dotted line is the cdf for the distances between individuals diagnosed with leukaemia between 1978 and 1982, whose addresses were known, for the eight county upstate New York area in Figure 3.

of the variance–covariance matrix of the differences [17, p. 44]. This formulation makes it clear that the geography of the region in question plays no role in such a statistic, since the statistic is invariant to permutations of the physical position of the locations, and thus is in general not likely to have good power against most alternative hypotheses of interest—in particular, clustering, which is a geographic phenomenon.

To overcome this shortcoming, Tango [18] proposes replacing the inverse of the variance–covariance matrix with one that reflects the distances between the locations. He defines a statistic T, in which he chooses to bring the distances between individuals into play by using a weight function with weights exponentially decaying as the distance increases. Whittemore *et al.* [19] take a different tack. They argue that the fundamental variable of interest is the distances between individuals, and consider the average (δ) of these distances. While we agree with the authors that consideration of the distances between individuals is pivotal to this problem, we feel that averaging may be too severe a summarization. This feeling is borne out by the power study in Section 4.

3.2. Leukaemia in upstate New York

Figure 1 shows the cdf $F(\cdot)=F(\cdot; p)$ for a population of a little over 1 million individuals reported in the 1980 census in 790 census subdivisions defined over these 8 counties in upstate New York (shown in Figure 3 [left]). Also shown in Figure 1 is the ecdf $F_n(\cdot)=F(\cdot; \hat{p})$ of the distances between 581 individuals diagnosed with leukaemia during the 5-year period 1978–1982 in the region. (The real number of cases during that period of time was 592, but here we report only those whose location is known with certainty). The question of interest is whether the leukaemia cases in upstate New York show any evidence of geographic clustering over and above the natural clustering levels existing at the population centres, and if that is the case, where does the clustering occur. These data originated from the New York State Cancer Registry, and this example was first discussed in Reference [20], and later in Reference [21].

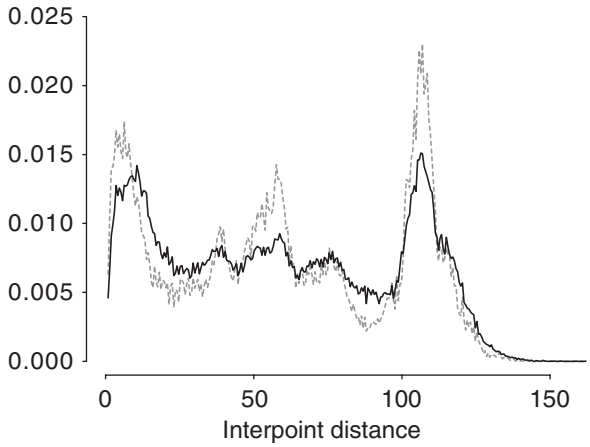

Figure 2. First difference functions of the two cdf curves in Figure 1. The solid line corresponds to the population density (the solid cdf) and the dotted line to the leukaemia density (the dotted cdf).

These authors apply various methods of analysis to this data, and we refer to those references for a description and comparison of those methods.

One can see a difference between the two functions displayed in Figure 1, but Figure 2 is much more visually satisfying. In the latter figure we show the two 'density' functions corresponding to the cdfs in Figure 1. For the density functions we used a grid of 300 equally spaced points.

One can distinguish between two kinds of clusters; we may call them *endogenous* and *exogenous*. An endogenous cluster is one apparent in the population distribution (such as a population centre) evidenced on the interpoint density function through the presence of peaks, as in the solid curve in Figure 2. For example, the peak around 110 km is mostly due to the clustering in the two major urban centres (Binghamton and Syracuse), while that around 60 km is mostly due to the population clustering in Binghamton and the other three more populated areas (from left to right, Cortland, Ithaca and Norwich) in the middle of the region, as well as the clustering in Syracuse and Cortland, since these five pairs of population centres are each approximately 60 km apart. Note also the smaller peaks at 40 km (distance from Syracuse to Auburn) and at 80 km (distance from Syracuse to both Ithaca and Norwich).

An exogenous cluster is one that is superimposed on the population distribution (with its existing endogenous clusters) and it is introduced by some force not uniformly evident in the whole population. The endogenous clusters are important because they form the baseline against which clusters need to be evaluated. In this application, we might suspect that the difference between $f_n(\cdot)$ and $f(\cdot)$ is big, and possibly too big to attribute to sampling variability, especially for small distances, as pointed out by a number of authors (see References [19, 22], for example). We contend that additional information is available in the discrepancy for larger distances as well, and that if we do not consider them, we are discarding power unnecessarily. Indeed, we see that there is an increase in the peaks near the origin, at about 60 km, at about 110 km and possibly even at 40 km in $f_n(\cdot)$ when compared to $f(\cdot)$, but that there is no increase at 80 km. Note that since the integrals under these two functions are the same,

the troughs must compensate for the excesses in the peaks. These effects help in identifying possible exogenous clusters. In fact, the big increases at 60 and 110 km can be explained rather nicely by a cluster of leukaemia cases around Binghamton. This would cause an increase in the frequency at the distances between Binghamton and the sites located at roughly 60 km (Ithaca, Cortland and Norwich) and 110 km (Syracuse) from Binghamton as is evident in the figure. Further, the lack of an increase at 80 km would indicate the lack of a cluster near Syracuse, Ithaca or Norwich. On the other hand, an increase in the frequency at 40 km can be caused by a cluster at Auburn (an increase at Syracuse would have also produced a peak at 80 km, and that peak was not observed). Note how these observations should be attempted only once the test statistic rejects the null hypothesis, as peaks and valleys will also occur under the null, and there would be risk of over-interpretation otherwise.

Testing for clustering using the proposed statistic M rejects the null hypothesis, at the 5 per cent level, that the leukaemia cases can be considered a random sample from these population centres ($p = 0.000$). When applying other existing statistics to this data set (see Section 4 below) we obtain p-values of 0.000 for T, 0.944 for DC, and 0.804 for δ implying that Tango's statistic is significant, but Diggle's and Whittemore's statistics do not find any evidence for clustering.

3.3. *Locating clusters*

Deciding that a sample exhibits evidence of clustering may not be an end unto itself, unless for example one is interested in establishing whether a disease is infectious. Typically, one is interested in the location(s) where the clustering may be occurring. A cluster will not only have an impact at a primary location (as exhibited by the behaviour of $f(\cdot)$ near the origin), but will also have secondary impacts on the peaks of $f(\cdot)$ at those distances that reflect its distances from other underlying clusters; typically, dense, urban areas. To locate where the disease-induced clusters may be in the discrete setting, we consider an (admittedly *ad hoc*) method based on decomposing the M statistic. We first decompose M to assign to each location its contribution to the total. To this end we rewrite M as

$$\begin{aligned} M(f_n(\mathbf{d}), f(\mathbf{d})) &= (f_n(\mathbf{d}) - f(\mathbf{d}))^t S^- (f_n(\mathbf{d}) - f(\mathbf{d})) \\ &= \sum_{h=1}^{m} (f_n(d_h) - f(d_h)) \sum_{t=1}^{m} s^{ht} (f_n(d_t) - f(d_t)) = \sum_{h=1}^{m} \Delta_h W_h \end{aligned}$$

where $\Delta_h = (f_n(d_h) - f(d_h))$ and W_h is the internal summation. From the definitions of $f(\mathbf{d})$ and $f_n(\mathbf{d})$, the contribution $\Delta_h W_h$ of each interval $(d_h - \varepsilon/2, d_h + \varepsilon/2]$ to M can be decomposed among each of the contributing pairs of locations (l_i, l_j), $i, j = 1, \ldots, k$ as

$$\Delta_h W_h = \sum_{i=1}^{k} \sum_{j=1}^{k} \Delta_h(i,j) W_h$$

with

$$\Delta_h(i,j) = \frac{1}{\varepsilon} 1(d_h - \varepsilon/2 < d_{ij} \leqslant d_h + \varepsilon/2) \left(\frac{n_i n_j}{n^2} - \frac{N_i N_j}{N^2} \right)$$

This contribution to the statistic M, $\Delta_h(i,j)W_h$, represents a contribution from two locations, l_i and l_j. How to make the attribution to each of these locations is not unique. We choose to consider the deviation between the observed proportions ($\hat{p}_i = n_i/n$) and the expected proportions ($p_i = N_i/N$) at those locations. To this end define,

$$\alpha(i,j) = \frac{|\hat{p}_i - p_i|}{|\hat{p}_i - p_i| + |\hat{p}_j - p_j|}$$

and assign $\alpha(i,j)\Delta_h(i,j)W_h$ and $(1-\alpha(i,j))\Delta_h(i,j)W_h$ to l_i and l_j, respectively. For each of the intervals $(d_h - \varepsilon/2, d_h + \varepsilon/2]$ for $h = 1,2,\ldots,m$, one can then define for each location l_i a total contribution to M (or 'score') equal to

$$\text{Score}(i) = \sum_{h=1}^{m} \sum_{j=1}^{k} \alpha(i,j)\Delta_h(i,j)W_h$$

It is easy to verify that $\sum_{i=1}^{k} \text{Score}(i) = M$, so that the scores decompose M. Note how this decomposition approach is similar in spirit to the examination of local statistics in the analysis of spatial autocorrelation (see References [23–25]).

The locations can then be ranked according to their score. In a particular data set, if the M statistic is significantly different from what would be expected under the null hypothesis, then the locations can be studied to see which locations impact M the most. One strategy for identifying locations with large contributions to M may be to consider the difference between the observed value of M and the cut-off M^* corresponding to the test, and find the minimum number of locations (having the largest scores) such that the sum of their scores equals $M - M^*$.

Application of this procedure yields the map on the right in Figure 3. In the figure, we highlight the top 13 locations selected. Even though the interpretation of the results of the cluster localization procedure is perhaps a bit beyond the scope of the proposed tests (and should therefore be taken with caution), the locations selected can be seen to be suspiciously close to some of the waste sites shown on the map. The number of locations to plot was chosen based on the fact that the difference between the observed value of M (144.1) and the cut-off point for the corresponding 5 per cent sampling test (44.6) is roughly equal to the sum of the scores of the top 13 locations (99.7). All of these locations show an excess in the number of leukaemia cases. To ensure stability of the estimated distribution of M we have used 32 bins for the calculation of the p-value and for the identification of suspicious locations. However, a p-value equal to zero and a figure very similar to Figure 3 were obtained when using 300 bins in the definition of M.

Consistent with the impression gained by contrasting $f_n(\cdot)$ with $f(\cdot)$, there is an indication that the locations around Binghamton form a cluster of locations with excess numbers of leukaemia cases. The locations so identified follow the flow of the Susquehanna river through that region. Two other areas identified are in the upper-left corner and in the middle of the map. These regions were also identified in Reference [20] using the geographical analysis machine method [26] designed for finding areas with high rates. Unfortunately, the latter method does not lead to a quantitative assessment of the significance of the observed pattern, so that it is hard to interpret its results. The possibility of clustering of cases around Binghamton was also indicated in Reference [27], where the hypothesis of randomness was also rejected. Their likelihood-based approach is constructed on the alternative 'hot-spot' model defined in

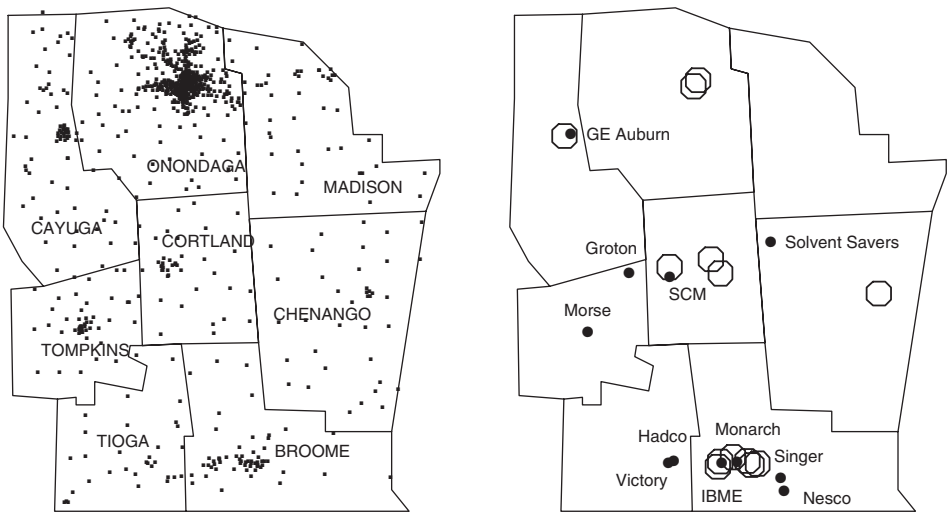

Figure 3. On the left is the map of the New York study area with location of cell centroids indicated by '·'. The dense cluster in Cayuga county is Auburn, the one in Onondaga is Syracuse, the one in Tompkins county is Ithaca, the one in Cortland county is Cortland, the one in Chenango is Norwich, and the one in Broome county is Binghamton. The graph on the right is the same region, with the identified 18 locations with large contributions to M shown by circles, and the labelled hazardous waste sites containing trichloroethylene indicated by dots.

Reference [28]; i.e. that the probability of leukaemia is elevated and constant within a particular radius of a point defined to be the centre of the cluster. We should note that the M statistic does not define an alternative hypothesis, but that this does not mean that it is good (or bad) for all alternative hypotheses, nor that methods based on probability models can only perform well only under those specific models.

Also from Figure 3, we see that the clusters around Cortland and Auburn are close to identified waste sites. The other three locations, one in Chenango and two in Onondago, are rather distant from all waste sites. Of course, these implied relationships are quite suggestive, but before one can make any more definitive statement one would need to investigate them further. In particular, the many issues associated with the study of the effects of exposure to toxic substances (whose quantity and toxicity should in general be expected to vary over the exposure period) are well beyond the scope of our work here. The migration patterns of the population across the region in the time period considered and any cumulative effect of exposure to the toxic substances (as well as the kinds of toxic substances) should all be considered before drawing any conclusions about the effect of the toxic waste sites on the population. Our methods do not attempt to solve such a complex and general problem, but rather our inference is limited to the study of deviations of the spatial distribution of the leukaemia cases from the underlying population distribution. As a consequence, Figure 3 should only be meant as a visual exploratory analysis of the possible connection between the locations of the sites and the distribution of the cases.

It is of interest to note that only two of the many locations in and around Syracuse (427 locations within a radius of 20 km from the centre of Syracuse) are identified as having excessive numbers of individuals with leukaemia, even though over 40 per cent of the region's population lives there. Thus the proposed method seems to show considerable specificity in this example.

4. A SIMULATION STUDY

We performed a power study to compare the proposed statistics M, M_1, M_{χ^2}, and M_{KL} defined in Section 2.3 with three well-known currently available and easily implementable alternative statistics: the δ statistic [19], the T statistic [18], and the DC statistic [22] based on Ripley's K-functions. The statistic DC was designed under the assumption of a Cox process, i.e. that the underlying distribution be a realization of a Poisson point process having as intensity the realization of a further probability distribution. This implies that the n_i should be zeroes or ones, but this constraint is usually ignored in application, and we continue in the same vein to use DC both in the discrete and in the continuous setting.

Whittemore and colleagues [19] derive the first two moments of the δ statistic and prove its asymptotic normality, but rather than rely on this asymptotic result we sample from the exact distribution since this would yield more accurate results. We do the same for the T statistic. For the DC statistic we use a ratio of 2 to 1 for the number of controls to the number of cases.

We consider two settings: first, the situation where points are distributed uniformly over the unit square; and second, the common situation of fixed locations over a highly non-homogeneous map (the New York State map described in Section 3.2) with more than one individual at each location.

4.1. Continuous homogeneous setting

We test the performance of the statistics under the homogeneous point process setting first proposed in Reference [29], and also discussed in Reference [22]. We follow the instructions for the simulation in Reference [22], as best we can, to generate the powers for the other statistics, and quote these authors for the power results of their statistic (their estimates are based on 100 simulations, ours on 1000). Under the null distribution a sample of $n_1 = 50$ points is generated uniformly on the unit square, while under the various alternatives (identified by the parameters q, v, and σ) some $50qv$ of these points are deleted and replaced by $50q$ clusters of v cases, with centres distributed completely at random and cluster members displaced independently from their corresponding cluster centre according to an isotropic bivariate normal distribution with standard deviation σ in either co-ordinate direction—thus with probability one no two points fall in the same location. We computed the power for δ and M (with $m = 20$, see Section 2.3) under some of the parameter combinations reported in Reference [22]. For the remaining parameter combinations we could not reconstruct the exact algorithm used to generate the samples as reported in that article since $50qv$ is not an integer. Note that Tango's T cannot be used immediately in this setting, so that it does not appear in the table below.

Table I. Estimates of power under the point process setting.

	q : 0.10			0.20			0.10		
	$v=2$						$v=4$		
$n_2=50$	δ	M	DC	δ	M	DC	δ	M	DC
$\sigma=0.001$	0.09	0.48	0.49	0.17	0.97	0.95	0.30	1.00	1.00
$\sigma=0.005$	0.12	0.43	0.40	0.13	0.96	0.90	0.29	1.00	1.00
$\sigma=0.01$	0.11	0.44	0.28	0.16	0.96	0.79	0.30	1.00	1.00
$n_2=200$	q : 0.02			0.04			0.06		
	$v=4$								
	δ	M	DC	δ	M	DC	δ	M	DC
$\sigma=0.01$	0.10	0.24	0.57	0.14	0.63	0.98	0.21	0.90	1.00

The entries for DC are quoted from Diggle and Chetwynd [23].

For the values of $q=0.1$ and 0.2 the DC statistic was based on 50 cases and 50 controls, while for the values of $q=0.02, 0.04$, and 0.06 it was based on 50 cases and 200 controls. The results from these power estimates are shown in Table I, and they show that the DC and M statistics should be preferred to δ in such a homogeneous setting, with the DC doing considerably better than M for smaller q, and M doing slightly better than DC in the case of several relatively large clusters ($q \geqslant 0.10$ and $\sigma=0.01$) and fewer controls. DC was constructed using 100 bins, but [22] also reports some results obtained using 10 bins and $v=2$. With that implementation of DC the performance of that statistic seems to improve for smaller σ, but deteriorates for larger σ.

4.2. Discrete inhomogeneous setting

For this first part of the power study we use the New York State population distribution described in Section 3.2. We construct the null distribution of the statistics to be studied by taking samples from the 790 census subdivisions' centroids with probabilities proportional to each subdivision's population count. We first consider samples of size 105 cases, and then 528 cases. These correspond to prevalences of 0.0001 and 0.0005, respectively. By sampling from these null hypotheses we establish the cut-off values for the Monte Carlo tests for the statistics being compared. The cut-off values are chosen to achieve a type I error level of 5 per cent.

We construct the alternative hypotheses by adding one cluster, placed at different locations to study the effect of the geography. To determine the placements, we sort the locations by the population density around them. This is done by computing the total number of individuals living within a circle of radius 10 km from each location. We then pick as a centre of the cluster for the alternative hypotheses in turn the locations corresponding to several percentiles of such a population density distribution. We call these locations Q10, Q15, Q20, Q25, Q30, Q40, Q50, and Q100, respectively, naming them after their corresponding percentiles.

All *deciles* between the median and the largest value correspond to locations within or around Syracuse, and they yield results similar to Q100 (that we label 'C' in Table II). Since we want a broader representation, we also hand-pick two more locations as positions for the cluster centre. These locations are in the middle of Auburn ('A') and Binghamton ('B') respectively, chosen as representatives of small and medium-sized urban centres. Binghamton is also chosen because of the interest in the potentially hazardous waste sites near that city. We saw in the previous section that the region around Binghamton is identified as a possible location of a cluster of leukaemia cases.

To study the extent of the influence of a cluster, a radius, ρ, around the cluster centre is chosen within which the probability of becoming diseased is elevated. We choose three values: $\rho = 2$, 5, and 10 km to indicate clusters with increasing impact. Within the radius of influence, we choose a factor κ by which to increase the probability of becoming diseased. At the centre of the cluster, the probability of becoming diseased is multiplied by $(1 + \kappa)$, and the increase-factor decreases linearly to one at the perimeter of the circle of radius ρ. (The probabilities are re-scaled to add to one). We choose different values κ, as shown in Table I. This is an example of a 'clinal' (or 'conic') cluster as defined in Reference [28].

We also study 'cylindric' clusters, i.e. clusters for which the same factor $(1 + \kappa)$ is applied to all locations falling within the cluster, irrespectively of their distance from the centre of the cluster. Among cylindric clusters we experiment with elliptically shaped clusters with ratios between the longest and the shortest diameter in turn equal to 1, 2.5, and 5. These clusters all have their smallest diameter equal to 4 km, so that they are uniquely identified as having ρ equal to 2, 5, and 10 km, respectively.

The powers of the statistics are estimated by counting the proportion of the samples (generated according to some alternative hypothesis) that are more extreme than the 5 per cent cut-off values obtained from the null distribution. The way in which we create the alternative hypotheses is such that putting a cluster on a densely populated area will have a stronger impact on the overall distribution of the cases than a cluster placed on an area of low population density, since we condition on the total number of cases. This way of creating alternative hypotheses thus makes clusters placed in highly populated areas easier to detect, and gives an overall impression of varying prevalence.

Table II shows that the power of all statistics varies with the location of the cluster centre, its extent (ρ), and the overall prevalence. The power of any statistic in general depends very strongly on the underlying population distribution as well as on all these parameters, but it seems clear that the proposed statistics M_1, M_{χ^2}, M_{KL} and M perform very well, and that in particular the power gain of M over all the other statistics is large. This is probably due to the fact that M is the only one among these statistics that explicitly accounts for the covariance structure in $f_n(\mathbf{d})$. Tango's T statistic performs quite well, especially when the cluster is placed in highly populated areas such as Q50 and B (in which cases it sometimes even outperforms all other statistics). Quite often, however, its power is much smaller than M's. Notice that we choose the parameter τ in the expression of T to be equal to 5, thus making beneficial use of prior information (external to the data) about the alternative hypotheses. That information is not usually available. In fact, expanding $\exp(-d/5)$ to the linear term gives $1 - d/5$, so that T gives most weight to deviations from the expected counts occurring in the same direction at pairs of locations that are roughly within 5 km of each other. Note that other weight matrices could be defined, that take into consideration an assumed spatial structure (see for example Reference [30]).

Table II. Results of power estimation. A is Syracuse, B is Binghamton, and C is Auburn (see text for additional definitions).

	$n = 105$										$n = 528$									
Location	Q10	Q15	Q20	Q25	Q30	Q40	Q50	A	B	C	Q10	Q15	Q20	Q25	Q30	Q40	Q50	A	B	C
Conic cluster																				
$\rho = 2,\ \kappa = 4^*$																				
M	**0.06**	**0.07**	**0.09**	**0.09**	**0.08**	**0.12**	0.07	**0.08**	**0.11**	**0.31**	**0.06**	**0.09**	**0.10**	**0.29**	**0.14**	**0.39**	0.18	**0.23**	**0.27**	**0.86**
M_1	0.05	0.03	0.04	0.04	0.05	0.07	0.06	0.04	0.08	0.21	0.05	0.08	0.06	0.09	0.12	0.37	0.15	0.14	0.16	0.61
M_{χ^2}	0.05	0.03	0.04	0.05	0.05	0.07	0.06	0.04	0.07	0.18	0.05	**0.09**	0.07	0.10	0.12	**0.39**	0.13	0.15	0.15	0.59
M_{KL}	**0.06**	0.06	0.06	0.06	0.06	0.07	**0.08**	0.06	0.09	0.19	**0.06**	**0.09**	0.07	0.09	0.12	0.37	0.14	0.14	0.15	0.54
T	0.05	0.06	0.05	0.06	0.06	0.06	0.07	0.05	0.07	0.07	**0.06**	0.06	0.05	0.08	0.05	0.16	**0.22**	0.08	0.21	0.23
δ	0.05	0.06	0.06	0.07	0.05	0.04	0.06	0.06	0.06	0.11	0.05	0.06	0.07	0.08	0.05	0.05	0.14	0.07	0.14	0.36
DC	**0.06**	0.06	0.05	0.06	0.05	0.04	0.05	0.06	0.05	0.10	**0.06**	0.05	0.06	0.06	0.05	0.05	0.10	0.09	0.09	0.32
$\rho = 5,\ \kappa = 2^*$																				
M	**0.06**	**0.07**	**0.08**	**0.13**	**0.08**	**0.15**	0.11	**0.10**	**0.23**	**0.67**	**0.06**	**0.09**	**0.15**	**0.66**	0.19	0.61	0.34	**0.42**	**0.80**	**1.00**
M_1	0.05	0.03	0.04	0.05	0.06	0.12	0.08	0.07	0.18	0.61	0.05	0.08	0.08	0.24	0.20	0.62	0.28	0.28	0.70	**1.00**
M_{χ^2}	0.05	0.03	0.04	0.05	0.06	0.12	0.08	0.07	0.16	0.57	0.05	**0.09**	0.10	0.25	**0.22**	**0.64**	0.25	0.29	0.65	**1.00**
M_{KL}	**0.06**	0.06	0.07	0.06	**0.08**	0.13	0.09	0.08	0.16	0.51	**0.06**	**0.09**	0.09	0.21	0.20	0.63	0.25	0.25	0.62	0.99
T	0.05	0.06	0.06	0.07	0.05	0.08	**0.12**	0.05	0.21	0.21	**0.06**	0.06	0.06	0.23	0.06	0.33	**0.43**	0.15	0.73	0.82
δ	0.05	0.06	0.06	0.09	0.06	0.04	0.07	0.07	0.12	0.31	0.05	0.06	0.09	0.10	0.08	0.04	0.22	0.08	0.46	0.91
DC	**0.06**	0.06	0.05	0.06	0.05	0.05	0.05	0.07	0.06	0.27	**0.06**	0.05	0.07	0.07	0.05	0.05	0.16	0.12	0.25	0.85
$\rho = 10,\ \kappa = 1^*$																				
M	**0.07**	0.04	**0.08**	**0.30**	**0.08**	**0.10**	0.07	**0.08**	**0.16**	0.43	0.07	0.11	**0.52**	**0.97**	0.14	0.28	0.17	**0.18**	0.52	0.97
M_1	0.03	0.05	0.06	0.10	0.05	0.07	0.08	0.05	0.13	**0.51**	0.06	**0.19**	0.37	0.87	0.18	0.35	0.22	0.12	0.51	**0.98**
M_{χ^2}	0.04	0.05	0.07	0.12	0.04	0.06	0.06	0.05	0.13	0.45	0.07	**0.19**	0.44	0.87	**0.19**	**0.37**	0.21	0.13	0.47	0.97
M_{KL}	0.05	**0.06**	**0.08**	0.13	0.07	0.08	0.09	0.07	0.14	0.43	0.07	**0.19**	0.41	0.84	0.18	0.35	0.20	0.14	0.45	0.97
T	0.06	0.05	0.06	0.23	0.06	0.06	**0.11**	0.06	0.18	0.22	0.07	0.07	0.14	**0.97**	0.06	0.15	**0.37**	0.08	**0.66**	0.80
δ	**0.07**	0.04	**0.08**	0.10	0.05	0.04	0.06	0.06	0.10	0.31	**0.08**	0.09	0.19	0.11	0.08	0.05	0.20	0.07	0.40	0.91
DC	0.06	0.05	0.06	0.07	0.05	0.05	0.06	0.06	0.06	0.24	**0.08**	0.05	0.09	0.10	0.05	0.05	0.14	0.09	0.23	0.82

Table II. *Continued.*

	$n = 105$										$n = 528$									
Location	Q10	Q15	Q20	Q25	Q30	Q40	Q50	A	B	C	Q10	Q15	Q20	Q25	Q30	Q40	Q50	A	B	C
Cylindric cluster																				
$\rho = 2,\ \kappa = 4^*$																				
M	**0.06**	**0.07**	**0.09**	**0.09**	**0.09**	**0.44**	**0.21**	**0.32**	**0.32**	**0.88**	**0.06**	**0.09**	**0.10**	**0.39**	**0.28**	**0.99**	0.73	**0.96**	**0.91**	**1.00**
M_l	0.05	0.03	0.04	0.05	0.06	0.34	0.16	0.20	0.21	0.76	0.05	0.08	0.06	0.12	0.24	0.98	0.60	0.92	0.79	**1.00**
M_{χ^2}	0.05	0.03	0.04	0.04	0.06	0.35	0.14	0.19	0.20	0.74	0.05	**0.09**	0.07	0.13	0.25	**0.99**	0.55	0.92	0.76	**1.00**
M_{KL}	**0.06**	0.06	0.06	0.05	0.08	0.34	0.17	0.20	0.20	0.65	**0.06**	**0.09**	0.07	0.12	0.23	0.98	0.53	0.89	0.73	**1.00**
T	0.05	0.06	0.05	0.05	0.06	0.19	**0.21**	0.11	0.23	0.26	**0.06**	0.06	0.05	0.12	0.06	0.90	**0.75**	0.68	0.79	0.92
δ	0.05	0.06	0.06	0.07	0.06	0.03	0.10	0.06	0.12	0.37	0.05	0.06	0.07	0.10	0.08	0.04	0.41	0.13	0.47	0.95
DC	**0.06**	0.06	0.05	0.05	0.05	0.03	0.07	0.09	0.07	0.35	**0.06**	0.05	0.06	0.06	0.05	0.04	0.25	0.25	0.24	0.93
$\rho = 5,\ \kappa = 2^*$																				
M	**0.06**	**0.07**	**0.10**	**0.50**	**0.11**	**0.25**	0.20	**0.18**	**0.55**	**0.95**	**0.06**	**0.09**	**0.23**	**1.00**	0.39	0.91	0.80	**0.82**	**1.00**	**1.00**
M_l	0.05	0.03	0.04	0.17	0.09	0.20	0.17	0.11	0.46	0.93	0.05	0.08	0.12	0.93	0.44	0.91	0.70	0.70	0.99	**1.00**
M_{χ^2}	0.05	0.03	0.04	0.18	0.08	0.20	0.16	0.11	0.44	0.92	0.05	**0.09**	0.14	0.93	**0.45**	**0.92**	0.65	0.70	0.99	**1.00**
M_{KL}	**0.06**	0.06	0.06	0.17	0.09	0.20	0.18	0.12	0.41	0.88	**0.06**	**0.09**	0.13	0.90	0.42	0.90	0.63	0.65	0.99	**1.00**
T	0.05	0.06	0.05	0.19	0.05	0.12	**0.26**	0.07	0.52	0.48	**0.06**	0.06	0.06	0.96	0.07	0.65	**0.86**	0.38	**1.00**	**1.00**
δ	0.05	0.06	0.06	0.12	0.05	0.04	0.09	0.07	0.13	0.64	0.05	0.06	0.10	0.15	0.10	0.04	0.46	0.10	0.85	**1.00**
DC	**0.06**	0.06	0.06	0.07	0.04	0.04	0.07	0.08	0.08	0.58	**0.06**	0.05	0.06	0.10	0.05	0.04	0.31	0.19	0.44	**1.00**
$\rho = 10,\ \kappa = 1^*$																				
M	**0.06**	**0.07**	**0.12**	0.56	**0.08**	**0.12**	0.12	**0.08**	0.24	0.60	**0.06**	**0.15**	**0.78**	**1.00**	0.19	0.42	0.43	**0.23**	0.81	**1.00**
M_l	0.05	0.05	0.06	0.22	0.06	0.09	0.14	0.06	0.23	**0.74**	0.05	0.12	0.44	**1.00**	0.24	0.54	0.57	0.17	0.86	**1.00**
M_{χ^2}	0.05	0.05	0.07	0.25	0.06	0.10	0.15	0.05	0.22	0.69	0.05	0.14	0.53	**1.00**	**0.26**	**0.58**	0.51	0.19	0.83	**1.00**
M_{KL}	**0.06**	**0.07**	0.08	0.26	**0.08**	0.11	0.17	0.07	0.23	0.65	0.06	0.14	0.48	**1.00**	0.25	0.54	0.50	0.18	0.81	**1.00**
T	0.05	0.06	0.06	**0.58**	0.05	0.07	**0.22**	0.05	**0.35**	0.33	**0.06**	0.06	0.14	**1.00**	0.06	0.28	**0.76**	0.10	**0.94**	0.97
δ	0.05	0.06	0.07	0.09	0.05	0.04	0.10	0.07	0.14	0.48	0.05	0.08	0.18	0.10	0.09	0.05	0.42	0.07	0.68	0.99
DC	**0.06**	0.05	0.06	0.11	0.04	0.04	0.07	0.06	0.09	0.39	**0.06**	0.05	0.09	0.19	0.05	0.04	0.29	0.09	0.41	0.96

*For locations Q10–Q25 the value $\kappa = 10$ was used throughout to ensure that the added cluster had a detectable impact.
Bold numbers indicate highest powers. Each estimate is based on 1000 replicates.

The performance of δ and DC in this setting is quite disappointing, with powers greater than 0.50 only when large clusters are placed at the two highly populated locations B or C. The power estimates for δ shown in Table II are based on the use of a two-sided test rather than on a one-sided test as may at first seem appropriate. The one-sided test (in the direction of rejecting the null hypothesis of randomness when δ is too small) may possibly work well for uniform underlying populations, but it creates problems for general populations, since the strong dependence among the interpoint distances can cause the statistic to actually be driven in the *opposite* direction as the intensity of an added cluster is increased. In fact, we also compute the powers corresponding to the one-sided test in the simulations (data not shown), and in several instances they result in powers for δ equal to zero because of this phenomenon.

The overall performance of the statistic M appears to be superior to that of δ, T, and DC, especially from the point of view of the robustness of their performance as the cluster is placed in different positions. Examination of Table II shows that these results are consistent across the two kinds of clusters (cylinder vs conic). However, care should be taken, as always, when interpreting any simulation results, because of their restricted generalizability.

5. DISCUSSION

We describe the use of the interpoint distribution function as a statistic for the description of spatial patterns, and in particular we use it to assist in the detection of clustering that may exist over and above the natural clustering present in the underlying population. Clearly, no simulation study can provide absolute conclusions about the properties of any of the statistics discussed here. From our experiment there is indication that the interpoint distance distribution methods perform well when the underlying population is highly inhomogeneous (although this is not necessarily the case in all applications, see for example Reference [31]). The interpoint distance distribution even seems to perform reasonably well when the points are generated according to a homogeneous distribution, but in that setting the DC statistic [22] performs better, especially when one uses a large number of controls in the computation of DC. We thus suggest that the M statistic should be added to the researcher's toolbox when assessing the possible presence of disease clusters over inhomogeneous populations.

On a more theoretical level, our M statistic shares some similarities with DC. The latter was designed for the setting in which no two points can share the same co-ordinates, as their approach extends the work of Ripley [5] to construct the statistic DC that is based on the difference between K-functions. The K-function resembles a little the ecdf of the distances between individuals, even though the former is unbounded. One shortcoming of the K-function is that it cannot be estimated with any degree of accuracy for distances beyond a small neighbourhood of each observation, and in fact it can be estimated only for distances up to half the maximal distances between the individuals on the map. This shortcoming implies that no information can be gained from larger interpoint distances, while the presence of a cluster may have a great impact on those distances, as is indeed the case in the example we present. The K-function approach seems designed to detect a clustering process (thought of as 'coagulation', meant as the process of creating many small clusters) rather than the addition of one (or a few) clusters to an existing population. In fact, the K-function is a second moment measure of the entire point process and, like a covariance, it is a summary of clustering/regularity behaviour over all observed events. A single, very localized cluster may not induce much evidence for clustering over the entire observed process.

In contrast to that approach, the interpoint distance distribution considered here is conditional on the region, and it summarizes the behaviour of the interpoint distance over its whole range and not only for smaller values. For example, in the New York state application, the largest distance between any two individuals is about 162 km, while the circumradius is about 80. This precludes the DC statistic from considering the quite informative peak at 110 km. In fact, our method is based on conditioning on the region actually observed—as opposed to trying to estimate the second-order characteristics of an underlying process, as the K-functions do, an undertaking of somewhat questionable value in the inhomogeneous setting. This may explain some of the superiority of the power characteristics of M for the alternatives considered in the application. Another difference between the two statistics is the consideration of the covariance structure of the cdf in the definition of M, which seems to be an effective way of capturing the strong dependence implicit in the very definition of interpoint distances. We believe that these differences explain the power observed for M in the simulation study, in particular in the New York State setting. On the other hand, when the underlying process is a homogeneous point process—i.e. when concentration on the interpoint distances close to zero is most informative—then the K-function approach seems to perform better than M in some cases. This could also be due to the absence of endogenous clusters.

Note also that for the very definition of a K-function there needs to be an underlying space on which one can define a (preferably homogeneous) point process, while there is no such requirement for the interpoint distance distribution; in the latter, the definition of a distance or dissimilarity measure suffices. The stated assumption of independence between the points does provide (in the continuous setting) the underpinnings for a Poisson approximation to the underlying spatial distribution as the number of points goes to infinity [32], but we feel that it is more natural not to rely on asymptotics (whose accuracy is questionable) but rather to work with the actual exact distributions whenever possible, as we have done here.

The lack of power of δ suggests that just considering the mean distance is perhaps too drastic a summary of the whole distribution of the interpoint distances. Tango's T statistic performed quite well under certain conditions, but not very well under others. Like DC, T also does not make full use of the information contained in the distribution of the interpoint distances at large distances, since it concerns itself with local behaviour. Also, the estimated powers for both δ and T do change quite a bit depending on whether the tests are one- or two-sided, highlighting the difficulties in the definitions and interpretation of these two statistics.

Tango [18] shows an interesting example of why he considers the δ statistic inappropriate for use over inhomogeneous populations. To whit, consider an artificial study area comprising of three locations in an equilateral triangle, and $p=(0.2,0.3,0.5)$. It is easy to show that δ takes on the same value both when there is no clustering and $\hat{p}=p$, and when there is clustering and $\hat{p}=(0.5,0.3,0.2)$ (a clear deviation from randomness). In this example all the interpoint distances are equal, so that δ is actually invariant to all of the 6 possible permutations of the elements of p. A similar argument can be made against the interpoint distance distribution. One cannot rule out the possibility that two different spatial distributions may yield the same $F(d)$. However, in the discrete setting this only seems possible if there exist locations having the same set of distances from *all* of the other locations, and this situation seems extremely hard to achieve when the geography is not trivial. In the continuous setting the construction of such an example seems even harder.

It should also be noted that a similar argument can be made against the Tango statistic. The definition of the matrix A in T is such that its being positive definite is not guaranteed, so that there exist situations in which T itself may be equal to zero while $\hat{p} \neq p$. Also, Kulldorff [33] shows an example of a clustering point process designed to cause DC to be identically equal to zero. In general, the derivation of the properties of the mapping from the data to the statistics used to test for clustering is a difficult problem, and, because of its importance, it deserves continued investigation.

Observe that as one moves from the discrete model to the continuous model, one can think of the position of the individuals as being measured with increasing precision, so that in many cases one can think of the discrete setting as being a discretization of an underlying continuous process. The issues associated with the convergence from the discrete setting to the continuous setting, as one increases the resolution of the data, is one that deserves further study.

Note that while inhomogeneous spatial processes are also being studied (see for example References [34–36]), one can in contrast summarize the interpoint distance distribution approach as being a conditional, non-parametric approach. The interpoint distance distribution is clearly a function of the distribution of the observations (and in particular, of the region being considered), so that in general it is hardly identifiable with a parametric form. The use of the interpoint distance distribution is very intuitive and similar in spirit to the use of the empirical cdf. Consideration of the interpoint distance distribution and of its empirical estimator $F_n(\cdot)$ can thus be regarded as an extension of the commonly used non-parametric approach for random samples, with the advantage that the use of the empirical cdf of multivariate co-ordinates (or equivalently, the estimation of the corresponding intensity functions) is hard to accomplish in high dimensional settings (see Reference [37] for related work in two dimensions in the uniform case), whereas the interpoint distance can always be defined and used whenever a metric between observations is available (see Reference [38] for an example using genetic distances).

APPENDIX: WEAK CONVERGENCE OF $\sqrt{n}(F_n(\cdot) - F(\cdot))$

Let $(S, \mathscr{S}, P)$ be a probability space, and let $\{X_1, \ldots, X_n\}$ be an i.i.d. sample from the distribution P. We consider the asymptotic properties of the stochastic process $U_n(d) = \binom{n}{2}^{-1} \sum_{1 \leqslant i_1 < i_2 \leqslant n} 1(d(X_{i_1}, X_{i_2}) \leqslant d)$, which is asymptotically equivalent to $F_n(d)$. In general, if $\mathscr{H}$ is a measurable VC-subgraph class of real symmetric functions $h \in \mathscr{H}$ on S^2 with an envelope H square integrable for P^2, P a probability measure on $(S, \mathscr{S})$, then,

$$\{\sqrt{n}(U_n(h) - P^2 h) : h \in \mathscr{H}\} \rightarrow_{\mathscr{L}} \{4G_P(Ph) : h \in \mathscr{H}\} \quad \text{in } l^\infty(\mathscr{H})$$

where $U_n(h) = \binom{n}{2}^{-1} \sum_{1 \leqslant i_1 < i_2 \leqslant n} h(X_{i_1}, X_{i_2})$, $n \geqslant 2$, $P^2 h = \int h d(P \times P)$ and $G_P(\cdot)$ is a Gaussian process evaluated at the values Ph, $h \in \mathscr{H}$ (with covariance function $\mathrm{cov}(G_P(Ph), G_P(Pg))$, for $(Ph)(x) = \int h(x,u) dP(u)$) (see Reference [10, Theorem 5.3.3], specializing to $n = 2$). In our setting we let S be some bounded region of the plane. Then the class of functions $\mathscr{H}$ is $\mathscr{H} = \{1_{\{d(X_1, X_2) \leqslant t\}},\ t \in [0, t_{\max}]\}$, and we can take as the square integrable envelope the function $H(X_1, X_2, t) \equiv 1\ \forall (X_1, X_2) \in S^2$ and $\forall t \in [0, t_{\max}]$, where $t_{\max}$ is the maximum interpoint distance that can be observed on S. In fact, H is measurable, everywhere finite and square integrable,

given the boundedness of S. Since the indicators above are real symmetric functions of X_1 and X_2, it remains to prove only that $\mathscr{H}$ is indeed a measurable VC-subgraph class of functions on S. This is indeed so since the graphs of the indicators $1_{\{d(X_1,X_2)\leqslant t\}}$ are ordered by inclusion and therefore they cannot shatter any set of two or more points. This proves the result.

Because of the very definition of a Gaussian process, this general result implies that for a fixed value d the empirical cdf $F_n(d)$ has $\sqrt{n}$-convergence to $\mathscr{E}(d(X_1,X_2)\leqslant d)=F(d)$. More generally, the weak convergence implies that the joint asymptotic distribution of the centred empirical cdf $\sqrt{n}(F_n(\mathbf{d})-F(\mathbf{d}))=\sqrt{n}(F_n(d_1)-F(d_1),\ldots,F_n(d_m)-F(d_m))$ computed at a finite number of fixed values $\mathbf{d}=(d_1,\ldots,d_m)$ is multivariate normal with covariance matrix Σ as described by the process above. In particular, for two distances d_a and d_b, the corresponding covariance term in Σ is

$$\sigma_{a,b}=4(E[1\{d(X_1,X_2)\leqslant d_a, d(X_1,X_3)\leqslant d_b)\}]-P(d(X_1,X_2)\leqslant d_a)P(d(X_1,X_2)\leqslant d_b))$$

Lastly, the asymptotic distribution of $\tilde{M}$ can easily be obtained. In fact, if we call Σ^- a generalized inverse of the matrix Σ, then by Theorem 25 in Reference [39, p. 69] it follows that:

$$n\tilde{M}(F_n(\mathbf{d}),F(\mathbf{d}))=n[F_n(\mathbf{d})-F(\mathbf{d})]'\Sigma^-[F_n(\mathbf{d})-F(\mathbf{d})]$$

has as asymptotic distribution a χ^2 distribution with degrees of freedom equal to $rank(\Sigma^-\Sigma)$. In fact, the necessary and sufficient condition $\Sigma\Sigma^-\Sigma\Sigma^-\Sigma=\Sigma\Sigma^-\Sigma$ is true by definition of generalized inverse.

ACKNOWLEDGEMENTS

This work was supported in part by National Institutes of Health Grants AI28076 (NIAID) and LM07677-01 (National Library of Medicine).

REFERENCES

1. Borel E. *Traité du Calcul des Probabilités et de ses Applications*, vol. I. Gauthier-Villars: Paris, 1925.
2. Bartlett MS. The spectral analysis of two-dimensional point processes. *Biometrika* 1964; **51**:299–311.
3. Silverman BW. Limit theorems for dissociated random variables. *Advances in Applied Probability* 1976; **8**: 806–819.
4. Sheng TK. The distance between two random points in plane regions. *Advances in Applied Probability* 1985; **17**:748–773.
5. Ripley BD. The second-order analysis of stationary point processes. *Journal of Applied Probability* 1976; **13**:255–266.
6. Silverman BW, Brown TC. Short distances, flat triangles and poisson limits. *Journal of Applied Probability* 1978; **15**:815–825.
7. Brown TC, Silverman BW. Rates of Poisson convergence for U-statistics. *Journal of Applied Probability* 1979; **16**:428–432.
8. Ripley BD. *Statistical Inference for Spatial Processes*. Cambridge University Press: Cambridge, 1988.
9. van der Vaart AW. *Asymptotic Statistics*. Cambridge University Press: Cambridge, 1998.
10. de la Peña VH, Giné E. *Decoupling*. Springer: Berlin, 1999.
11. Dwass M. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 1957; **28**:181–187.
12. Rao CR, Mitra SK. *Generalized Inverse of Matrices and its Applications*. Wiley: New York, 1971.
13. Anderson NH, Titterington DM. Some methods for investigating spatial clustering, with epidemiological applications. *Journal of the Royal Statistical Society, Series A, General* 1997; **160**:87–105.
14. Centers for Disease Control. Guidelines for investigating clusters of health events. *Morbidity and Mortality Weekly Report* 1990; **39**:RR–11.

15. Caldwell GG. Twenty-two years of cancer cluster investigations at the centers for disease control. *American Journal of Epidemiology* 1990; **132**:S43–S47.
16. Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel J-F, Bertollini R (eds). *Disease Mapping and Risk Assessment for Public Health*. Wiley: New York, 1999.
17. Agresti A. *Categorical Data Analysis*. Wiley: New York, 1990.
18. Tango T. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine* 1995; **14**:2323–2334.
19. Whittemore AS, Friend N, Brown BW, Holly EA. A test to detect clusters of disease (corr: V75 p. 396). *Biometrika* 1987; **74**:631–635.
20. Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC. Monitoring for clustering of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* 1990; **132**(Suppl.): S136–S143.
21. Waller LA, Turnbull BW, Clark LC, Nasca P. Spatial pattern analyses to detect rare disease clusters. *Case Studies in Biometry*. Wiley: New York, 1994; 3–23.
22. Diggle PJ, Chetwynd AG. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics* 1991; **47**:1155–1163.
23. Getis A, Ord JK. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 1992; **24**:189–206.
24. Anselin L. Local indicators of spatial association—LISA. *Geographical Analysis* 1995; **27**:93–115.
25. Ord JK, Getis A. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* 1995; **27**:286–306.
26. Openshaw S, Craft AW, Charlton M, Birch JM. Investigation of leukemia clusters by use of a geographical analysis machine. *The Lancet* 1988; **1**(8580):272–273.
27. Kulldorff M, Nagarwalla N. Spatial disease cluster: detection and inference. *Statistics in Medicine* 1955; **14**: 799–810.
28. Wartenberg D, Greenberg M. Detecting disease clusters: the importance of statistical power. *American Journal of Epidemiology* 1990; **132**:156–166.
29. Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations (with comments). *Journal of the Royal Statistical Society*, *Series B*, *Methodological* 1990; **52**:73–104.
30. Rogerson PA. The detection of clusters using a spatial variation of the chi-square goodness-of-fit statistic. *Geographical Analysis* 1999; **31**:130–147.
31. Kulldorff M, Tango T, Park PJ. Power comparisons for disease clustering tests. *Computational Statistics and Data Analysis* 2003; **42**:665–684.
32. Daley DJ, Vere-Jones D. *An Introduction to the Theory of Point Processes*. Springer: Berlin, 1988.
33. Kulldorff M. Statistical methods for spatial epidemiology: tests for randomness. In *GIS and Health in Europe*, Gatrell A, Loytonen M (eds). Taylor and Francis: London, 1998.
34. Ogata Y, Katsura K. Likelihood analysis of spatial inhomogeneity for marked point patterns. *Annals of the Institute of Statistical Mathematics* 1988; **40**:29–39.
35. Stoyan D, Stoyan H. Non-homogeneous gibbs process models for forestry—a case study. *Biometrical Journal* 1998; **40**:521–531.
36. Cressie NAC. *Statistics for Spatial Data*. Wiley: New York, 1991.
37. Zimmermann DL. A bivariate Cramer–von Mises-type of test for spatial randomness. *Applied Statistics* 1993; **42**:43–54.
38. Kowalski J, Pagano M, DeGruttola V. A nonparametric test of gene region heterogeneity associated with phenotype. *Journal of the American Statistical Association* 2002; **97**:398–408.
39. Searle SR. *Linear Models*. Wiley: New York, 1971.
40. Bonetti M. Geometric methods in data analysis. *Ph.D. Thesis*, University of Connecticut, Storrs, CT, 1996.

