Chapter 4

The Distribution of Interpoint Distances

Marco Bonetti, Laura Forsberg,* Al Ozonoff,* and Marcello Pagano**

4.1 Introduction

Health surveillance systems are designed to collect data continuously, analyze them, and report the results in order to prevent and control diseases. Such a system that concentrates on patients presenting at an emergency department of a hospital or group of hospitals, with certain syndromes associated with biological weapons, could prove useful as an early warning system of a bioterrorist attack. If on any particular day or sequence of days, the number exceeds a predetermined amount, an alarm may be raised. Since the number of patients arriving at a hospital may be modelled as a random process, this alarm threshold may be set according to the usual hypothesis testing paradigm which is concerned with the two errors: raising false alarms, and missing the raising of a warranted alarm.

Typically, the number of patients arriving at a hospital is influenced by such covariates as the season of the year, and the day of the week (see [1] and references therein, for example) and these are indeed important in order to place any set of numbers in their proper context, but in this chapter we focus on additional aspects of the data and assume a simple model for the arrival of patients. Consider the model where the number of patients is a Poisson random variable with mean λ, and suppose this represents the number of individuals we expect on any given day. To further simplify matters, suppose we wish to set up a system that raises the alarm if there are too many individuals observed on any particular day. So, if λ is sufficiently large to accurately use the normal approximation, one might raise the alarm if more than $\lambda + 1.645\sqrt{\lambda}$ patients arrive on any particular day. A one-sided alarm system like this runs a 5% false-positive rate. The associated power curve is easy to calculate.

Such a system would seem optimal in possible early detection of a disturbance that impacts uniformly across the whole area under surveillance. Alternatively, if the disturbance is due to a single emitting source (such as happened in the anthrax catastrophe in Sverdlovsk [2]) or a number of such sources, so that the disturbance to the system is geographically localized

*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston MA 02115.

(for example, a triangular plume downwind from the offending source, as in Sverdlovsk) then this geographical information, if available (such as in the addresses of the patients), should prove useful. Similarly, if the disturbance to the system is some contagious agent, then one would again expect some geographic clustering amongst the patients. Both these instances argue for a surveillance system that not only looks at the number of patients, but simultaneously looks at the location of where the patients were afflicted. Such systems are the motivation for this study, and we present some initial thoughts on the subject in this chapter.

We look at the distribution of distances between individuals as a summary of information on patient locations. In the cases we have investigated, we have found that this distribution does not seem to be affected by time or even season, so that it can form the basis for normalcy. Unfortunately this distribution is not easy to characterize because it differs for every different geographic distribution, and intuition is often foiled when attempting to estimate this function. We first look at some simple examples that make this point. Subsequently, we look at the empirical distribution of a statistic to measure the deviations from normalcy of such a distribution.

Consider the locations $X_1, X_2, \ldots$, of patients arriving at random and indexed by the order in which they arrive. Denote by $D(X_i, X_j)$ the geographical distance between individuals at X_i and X_j. Consider too the distribution function $F(d) = Pr(D(X_i, X_j) \leq d)$ for nonnegative d, and assume that F is independent of i and j and is constant over time. Suppose that we have a long history of steady state behavior of a hospital admission system so that we may estimate F by its empirical counterpart with confidence and act as if F is known, and equal to this estimate.

Now consider a disturbance to the system that may be reflected in the locations from which the patients come. For example, a point of toxic emissions might infect a neighborhood and result in a large increase of patients coming to the hospital from that particular neighborhood. We may phrase the problem by asking whether the distance distribution associated with the latest group of patients is given by F. We show in [3] how to test hypotheses about the distribution of distances with power to detect unusual clustering amongst patients.

In our current context, suppose we check on a daily basis not only whether there are too many patients, but also whether there is unusual clustering amongst the patients. Consider one particular day and denote by n the number of patients arriving on that day. Denote by F_n the empirical cumulative distribution function (ecdf) based on the n patients; i.e. if the patients are located at $X_1, \ldots, X_n$ and their interpoint distances are $D(X_i, X_j), i, j = 1, \ldots, N$, then

$$F_n(d) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} I(D(X_i, X_j) \leq d) \qquad \forall d \geq 0. \tag{4.1}$$

Define a statistic T_n to measure the distance between F and F_n – below we consider a number of these statistics. In order to test the hypothesis that $X_1, \ldots, X_n$ is a random sample from the steady-state distribution of patients arriving at the hospital, we need the null joint distribution of the couplet, (n, T_n). The rest of this chapter considers the interpoint distance distribution and how it can be used to detect deviations from the steady-state distribution.

4.2 The interpoint distance distribution

4.2.1 The Continuous Case

Consider a point process such that the observations can appear anywhere inside some bounded region, S. Let the distribution P of points in S be absolutely continuous, and define a non-negative function d of pairs of observations in this region. Henceforth we generically refer to such a function as a distance, even though in subsequent developments we do not make use of the triangle inequality a distance function must satisfy. The cdf $F(\cdot)$ of the interpoint distance D between two independent points selected according to P, is then $F(d) = \mathcal{E}1(d(X_1, X_2) \leq d)$, where $1(\cdot)$ is the indicator function and $\mathcal{E}$ denotes expectation with respect to the $P \times P$ distribution; thus, on average, $F(d)$ is the proportion of distances less than or equal to d.

As an example in which the spatial distribution is analytically known, consider the case of a mixture of K bivariate normal distributions f_i on the plane, i.e. let

$$f_i(x) = N_2\left(\begin{pmatrix} \mu_{1_i} \\ \mu_{2_i} \end{pmatrix}, \sigma^2 I_2\right) = N_2(\mu_i, \sigma^2 I_2).$$

It is easy to show that in this case the interpoint distance $Y = d(X_1, X_2)$ between two points randomly generated from $f(x) = \sum_{i=1}^{K} \pi_i f_i(x)$ is distributed as the square root of a mixture of chi square densities. If each individual has probability π_i of belonging to each of the distributions f_i, then the density $f_{Y^2}(\cdot)$ can be written as $f_{Y^2}(\cdot) = \sum_{i=1}^{K} \sum_{j=1}^{K} g(i, j)\pi_i \pi_j$, where $g(i, j)$ is the density function of the $\chi^2(2, \frac{1}{4\sigma^2}((\mu_{1_i} - \mu_{1_j})^2 + (\mu_{2_i} - \mu_{2_j})^2))$ distribution.

With an extension of the usual definition of empirical distribution for random samples we define the ecdf $F_n(.)$ of the interpoint distances associated with a random sample $X_1, \ldots, X_n$ as defined in Equation 4.1, above.

As an illustration, Figure 4.1 below shows the smoothed interpoint distance density function estimated on all the $\binom{n}{2}$ dependent distances obtained from $n = 100$ points generated from such a mixture of 3 bivariate normal distributions (top histogram), and the density function estimated on the 10,000 distances computed from 10,000 independent *pairs* of points from that same distribution (lower-left graph). The histogram and smooth density estimates illustrate the closeness of the interpoint distance distribution of the dependent distances to the empirical distribution of the interpoint distance between two randomly chosen points, and the closeness of the latter density function to the theoretical density $f_D(\cdot)$. The ecdf (see Equation 4.1) of the dependent interpoint distances among n points in the plane thus is a well-defined and behaved summary of a configuration of observations.

The definition, and use, of the interpoint distribution function given above does not require that the point process be stationary, but if it is, a number of theoretical results follow. On the plane, [10] reports the distribution of the interpoint distances for randomly distributed points on the unit square and on the unit circle (results originally due to [11]). The latter distribution can be shown to be equal to

$$f_D(d) = \frac{4d}{\pi}\left\{\cos^{-1}\frac{d}{2} - \frac{d}{2}\sqrt{1 - \frac{d^2}{4}}\right\}, \quad d \in [0, 2].$$

Bartlett in [10] 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.

For fixed d, $F_n(d)$ is a V-statistic (see for example [17, p. 172]). The scaled distribution of $F_n(d)$ computed at a finite set of values, $d_1, \ldots, d_m$, converges to a multivariate normal

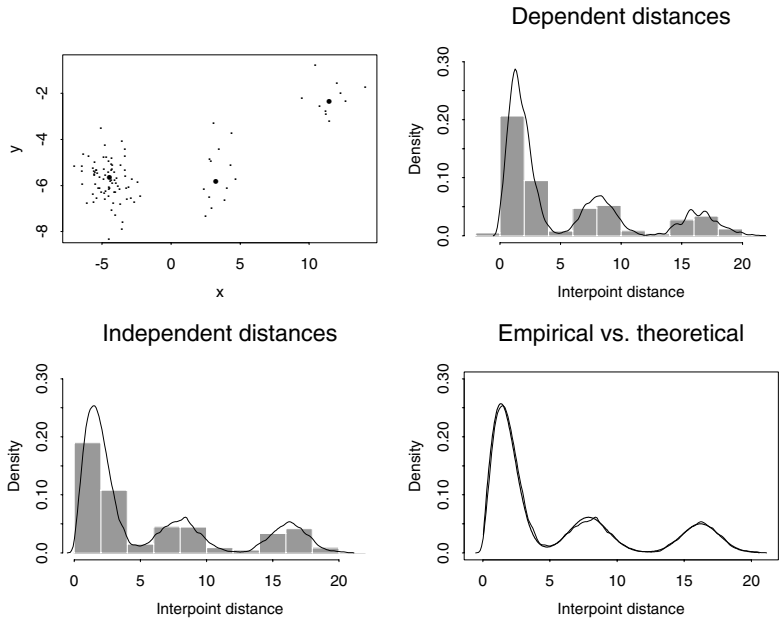

Figure 4.1. *An example of a realization of 100 points, top left, of a point process. The histogram from 10,000 dependent distances is given in the top right, and independent distances in the bottom left. The bottom right shows the comparison of these two smoothed estimates.*

distribution as $n \to \infty$. Also, the quantity $\sqrt{n}(F_n(d) - F(d))$, considered as a stochastic process indexed by d, converges weakly to a Gaussian process ([16], [3]). The use of this result, however, requires knowledge of the underlying spatial distribution, since the covariance function of the associated Gaussian process $GP(d)$ is equal to

$$\text{cov}\,(GP(d_1), GP(d_2)) = E\,[1(d(X_1, X_2) \le d_1)1(d(X_1, X_3) \le d_2)] - F(d_1)F(d_2).$$

If that is available, then the sampling distribution of the empirical interpoint distance distribution function can be obtained. Otherwise, it can be estimated via resampling methods or via simulation methods, as we do below. In lower dimensional settings one could estimate the intensity function via kernel methods, in which case as an alternative to the methods discussed here it is also possible to compare intensity functions ([14]).

4.2.2 The Discrete Case

Often continuous data is not available, so consider the case of a fixed population distribution with population centers $l_1, \ldots, l_k$ wherein live $N_1, \ldots, N_k$, individuals, respectively. For example, these may be the centers of census tracts or, on a smaller scale, houses. Let N be the total population size ($N = \sum_{i=1}^{k} N_i$). Let the random variable D represent the distance between two individuals chosen at random (with replacement) from this population. Formally, let $p_i = N_i/N,\ i = 1, \ldots k$ and $p = (p_1, \ldots, p_k)$, and let $N \to \infty$. Let d_{ij} be the distance between locations l_i and l_j. The random variable D then takes on the value d_{ij} with probability $p_i p_j$. The distribution function of this non-negative random variable is

thus,

$$F(d) = F(d;\, p) = \sum_{i=1}^{k} \sum_{j=1}^{k} p_i\, p_j\, 1(d_{ij} \le d). \tag{4.2}$$

Consider a random sample $n_1, \ldots, n_k$ of individuals distributed over the same geographic 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 $F_n(d) = F(d;\, \widehat{p})$, where $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_k)$ and for $i = 1, \ldots, k$, $\widehat{p}_i = n_i/n$. These definitions of $F(d;\, p)$ and $F(d;\, \widehat{p})$ are the discrete analogues, and are equivalent to those of $F(d)$ and $F_n(d)$ given above for the continuous case.

If one is interested in the distribution of the distances between individuals, and does not wish to make assumptions or inference about the value of the sample size, n, one may condition on it, and then use the distribution of the distances obtained by choosing samples of size n at locations l_i with probabilities p_i, $i = 1, \ldots, k$ as the null distribution (see [13]). Then the null hypothesis of random sampling from the population distribution corresponds to the null 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 $n \to \infty$), then for any fixed real d, $F(d;\, \widehat{p})$ is a strongly consistent estimator of $F(d;\, p)$. Some measure of the difference between $F(d;\, \widehat{p})$ 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. For a fixed value d the empirical cdf $F(d;\, \widehat{p})$ has $\sqrt{n}$-convergence to $\mathcal{E}(d(X_1, X_2) \le d)$. Moreover, there is convergence to a multivariate normal distribution when one computes the cdf at the finite set of values $d_1, d_2, \ldots, d_m$, and the covariance structure of the limiting distribution can be expressed analytically ([3]).

4.2.3 Two discrete examples of interpoint distance distribution

Example 1: Two points

Let n individuals be assigned to either location $A = (0, 0)$ or to location $B = (1, 0)$ with probabilities p_A and $p_B = 1 - p_A$, respectively. Denote by n_A and n_B the number of individuals assigned to the two locations. The matrix of the n^2 interpoint distances between individuals thus contains only the two values zero and one. In particular, a total of $n_A^2 + n_B^2$ zero distances are observed, and a total of $2\, n_A\, n_B$ distances equal to one. The relative frequency distribution of the interpoint distances is thus defined by the two proportions $P_0 = (n_A^2 + n_B^2)/n^2$ and $P_1 = 2\, n_A\, n_B/n^2$. The expected values of P_0 and P_1 are equal to

$$\begin{aligned} E\,P_0 &= \frac{E n_A^2 + E n_B^2}{n^2} = \frac{(n\, p_A\, (1 - p_A) + n^2\, p_A^2) + (n\, p_B\, (1 - p_B) + n^2\, p_B^2)}{n^2} \\ &= \frac{2 p_A p_B}{n} + p_A^2 + p_B^2 \\ E\,P_1 &= \frac{2E(n_A n_B)}{n^2} = \frac{2E(n_A(n - n_A))}{n^2} = \frac{2(nE(n_A) - E(n_A^2))}{n^2} \\ &= 2 p_A p_B - \frac{2 p_A p_B}{n}, \end{aligned}$$

since n_A and n_B are binomial random variables. Note how the expected value of P_1 follows immediately from that of P_0, as $P_1 = 1 - P_0$. As n tend to infinity the expected

values converge to the distribution of the interpoint distance $D = D(X_1, X_2)$ between two independent points X_1 and X_2, i.e., to the two probabilities $P(D = 0) = p_A^2 + p_B^2$ and $P(D = 1) = 2p_A p_B$.

The terms of the variance-covariance matrix of (P_0, P_1) for fixed n are equal to

$$\begin{aligned}
\text{var}(P_0) &= \frac{1}{n^4}\text{var}(n_A^2 + n_B^2) = \frac{1}{n^4}(En_A^4 + En_B^4 + En_A^2 n_B^2 - (En_A^2 + En_B^2)^2) \\
\text{var}(P_1) &= \text{var}(1 - P_0) = \text{var}(P_0) \\
\text{cov}(P_0, P_1) &= E(P_0 P_1) - E(P_0)E(P_1) = E(P_0(1 - P_0)) - E(P_0)E(1 - P_0) = -\text{var}(P_0)
\end{aligned}$$

From differentiation of the moment generating function of the binomial $\psi_{\mathbf{X}}(\mathbf{t}) = (p_A \exp(t_A) + p_B \exp(t_B))^n$ one obtains all necessary moments, and after some algebra the result

$$\begin{aligned}
\text{var}(P_0) = -\frac{4p_A}{n^3}\big(&6p_A^3 - 10np_A^3 + 4n^2 p_A^3 - 12p_A^2 - 12np_A + 7p_A - 1 \\
&+20np_A^2 - 8n^2 p_A^2 + 2n - n^2 + 5n^2 p_A\big) \\
\text{var}(P_1) &= \text{var}(P_0) \\
\text{cov}(P_0, P_1) &= -\text{var}(P_0)
\end{aligned}$$

Note how the last two expressions also follow immediately from the fact that $P_1 = 1 - P_0$. Thus we can focus on P_0 alone. For $p_A = p_B = 1/2$ the expected value of P_0 is equal to $(1 + 1/n)/2$, which tends to 1/2 as $n \to \infty$. Also, the expression for the variance of P_0 becomes

$$\text{var}(P_0) = \frac{1}{2}\frac{n-1}{n^3}.$$

Thus $\text{var}(\sqrt{n}P_0) \to 0$ as $n \to \infty$. Rescaling by n instead of $\sqrt{n}$ yields that

$$\text{var}\,(n\,P_0) \to \frac{1}{2} \text{ as } n \to \infty.$$

In fact, this happens if and only if $p_A = 1/2$, since in the expression of $\text{var}(P_0)$ one needs the condition $4\,p_A^3 - 8\,p_A^2 + 5\,p_A - 1 = 0$ to be satisfied, and $1/2$ is the only solution in $(0, 1)$. As a final remark, note how $1/2$ is *not* the variance of the binomial random variable with probability of succes (i.e., of falling into the value $D = 0$) equal to 1/2.

From the familiar results about U-statistics one might expect that $\sqrt{n}P_0$ has a non-degenerate asymptotic distribution, while from the expressions above it is clear that this does not happen. In fact, one would need to normalize P_0 by multiplication by n and not by $\sqrt{n}$ to converge to non-zero variances. The reason why this happens is worth discussing in detail. Below we refer for simplicity to one-dimensional U statistics, but with minor changes the same considerations apply to V-statistics such as the ones considered here.

This phenomenon is due to the fact that one of the requirements for the usual asymptotic normality results of U-statistics is not satisfied whenever $p_A = p_B$. In fact, consider the m-order U statistic

$$U_n = \frac{1}{n_{(m)}}\sum h(X_{i_1}, \ldots, X_{im})$$

with (symmetric) kernel $h(X_1, \ldots, X_m)$ (such that $E[h^2] < \infty$). The summation is taken over all $n_{(m)} = n(n-1)\cdots(n-m+1)$ m-tuples $(i_1, \ldots, i_m)$ of distinct elements from

$\{1, \ldots, n\}$. It is well known ([17] p.162) that if one defines the auxiliary functions $h_d(x_1, \ldots, x_d) = E[h(x_1, \ldots, x_d, X_{d+1}, \ldots, X_m)]$ and the parameters $\zeta_0 = 0$ and, for $1 \leq d \leq m$, $\zeta_d = \text{var}(h_d(X_1, \ldots, X_d))$, then the variance of U_n can be expressed as

$$\begin{aligned}\text{var}(U_n) &= \binom{n}{m}^{-1} \sum_{d=1}^{m} \binom{m}{d}\binom{n-m}{m-d}\zeta_d \\ &= \frac{m^2\zeta_1}{n} + O(n^{-2}).\end{aligned}$$

In our case above we have $m = 2$, $h(X_1, X_2) = 1(d(X_1, X_2) = d)$, for $d = 0$. (A symmetric argument with $d = 1$ would be used for X_1). Then one has

$$\begin{aligned}\zeta_1 &= \text{var}(h_1(X_1)) = \text{var}(E[h(x_1, X_2)]) = \\ &= \text{var}(P(d(x_1, X_2) = d))\end{aligned}$$

and

$$P(d(x_1, X_2) = 0) = \begin{cases} p_A & x_1 = (0, 0) \\ p_B & x_1 = (1, 0) \end{cases},$$

so that $p_A = p_B = 1/2 \Rightarrow \zeta_1 = 0$. In fact, here this is indeed a necessary and sufficient condition for the convergence in probability of $\sqrt{n}P_0$ to $P(D = 0) = 1/2$.

For a connection with related problems in genetics we refer the reader to [18].

This example is a degenerate one in the sense that one deals with only one (binomial) random variable. We now present another example that is constructed on four points and allows for the interpoint distance to take on four different values.

Example 2: Four points

Consider the situation of each of n individuals being assigned to one of the four points $A = (0, 2)$, $B = (1, 2)$, $C = (2, 0)$, and $D = (0, 0)$. The interpoint distance between two individuals can take on one of the four values $\{0, 1, 2, \sqrt{5}\}$. If one calls D_{ij} the interpoint distances D_{ij} between individuals i and j, one can then define the quantities $X_d = \sum_{i=1}^{n}\sum_{j=1}^{n} 1(D_{ij} = d)/n^2$, i.e. the proportion of interpoint distances equal to d. In particular,

$$\begin{aligned}P_0 &= \frac{1}{n^2}\left(n_A^2 + n_B^2 + n_C^2 + n_D^2\right) \\ P_1 &= \frac{2}{n^2}\left(n_A\, n_B + n_C\, n_D\right) \\ P_2 &= \frac{2}{n^2}\left(n_A\, n_D + n_B\, n_C\right) \\ P_{\sqrt{5}} &= \frac{2}{n^2}\left(n_A\, n_C + n_B\, n_D\right),\end{aligned}$$

where n_A, n_B, n_C, and n_D are the numbers of individuals at the four location, which we assume are assigned according to the multinomial distribution with parameters n and $\mathbf{p} = (p_A, p_B, p_C, p_D)$ for the four locations. Similarly to what was done in Example 1 one can

derive the expressions for the expected values and for the variance-covariance matrix of the vector $\mathbf{P} = [P_0, P_1, P_2, P_{\sqrt{5}}]$. The expressions for the expected values are as follows:

$$
\begin{aligned}
E\,P_0 &= (n\,p_A^2 + p_A - p_A^2 + n\,p_B^2 + p_B - p_B^2 + n\,p_C^2 + p_C - p_C^2 + n\,p_D^2 + p_D - p_D^2)/n \\
E\,P_1 &= 2(p_A\,p_B + p_D\,p_C)\frac{n-1}{n} \\
E\,P_2 &= 2(p_A\,p_D + p_B\,p_C)\frac{n-1}{n} \\
E\,P_{\sqrt{5}} &= 2(p_A\,p_C + p_B\,p_D)\frac{n-1}{n}
\end{aligned}
$$

The expression of the variance-covariance matrix for general $\mathbf{p}$ is quite messy.

As $n \to \infty$ the expected value of $\mathbf{P}$ converges to the vector of probabilities $P(D = d)$, where D is the distance between two individuals placed at random, i.e.,

$$
\begin{aligned}
P(D = 0) &= p_A^2 + p_B^2 + p_C^2 + p_D^2 \\
P(D = 1) &= 2(p_A\,p_B + p_C\,p_D) \\
P(D = 2) &= 2(p_A\,p_D + p_B\,p_C) \\
P(D = \sqrt{5}) &= 2(p_A\,p_C + p_B\,p_D).
\end{aligned}
$$

For the special case of $\mathbf{p} = [.25, .25, .25, .25]$ the expected value $E(\mathbf{P})$ becomes

$$
E\begin{bmatrix} P_0 \\ P_1 \\ P_2 \\ P_{\sqrt{5}} \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 1 + 1/n \\ 1 - 1/n \\ 1 - 1/n \\ 1 - 1/n \end{bmatrix},
$$

and the variance-covariance matrix takes the simple form

$$
\frac{1}{8}\left(\frac{1}{n^2} - \frac{1}{n^3}\right)\begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix}.
$$

Here, too, as $n \to \infty$ a phenomenon similar to that observed in Example 1 above occurs, namely the fact that $\text{var}(\sqrt{n}\mathbf{P}) \to \mathbf{c}$ with $\mathbf{c} = [.25, .25, .25, .25]'$, and that $\text{var}(n\mathbf{P}) \to [0, 0, 0, 0]'$.

4.2.4 Test statistics

It is easier and more informative to visualize the differences between these two cdfs $F_n(d)$ and $F(d)$ if we define the scaled first difference function $f(\mathbf{d})$. This is defined as a vector $f_n(\mathbf{d}) = (f_n(d_1), \ldots, f_n(d_m))$ of values

$$
f_n(d) = \frac{1}{\epsilon}\left[F_n(d + \epsilon/2) - F_n(d - \epsilon/2)\right]
$$

computed at the values $d_1, \ldots, d_m$ such that $d_j - d_{j-1} = \epsilon$ for $j = 1, \ldots, m$ and m some positive integer. We set $d_1 = \epsilon/2$, and for definiteness we define $f_n(d_1) = F_n(\epsilon)/\epsilon$

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 as $f_n(\mathbf{d})$ and at the same values $d_1, \ldots, d_m$, but replacing $F_n(\cdot)$ by $F(\cdot)$. (The constant ϵ may be any size and can be made as small as the accuracy with which the distances are measured.)

Because of its linear relationship with $F_n(\cdot)$, the first difference function $f_n(d)$ has $\sqrt{n}$-convergence to the expected value, $\mathcal{E}(1(d - \epsilon/2 < d(X_1, X_2) \leq d + \epsilon/2)$, and that for a fixed d, $n^{1/2} f(d; \widehat{p})$ has an asymptotically normal distribution. The joint asymptotic distribution for multiple values of d also follows immediately.

Several test statistics can be defined to measure the distance between $\widehat{F}_n(\cdot)$ and $F(\cdot)$, and thus allow the testing for deviations from the null spatial distribution. The asymptotic normality noted above suggests the use of the following statistic to measure the distance between the two vectors $f(\mathbf{d})$ and $f_n(\mathbf{d})$:

$$M(f_n(\mathbf{d}), f(\mathbf{d})) = (f_n(\mathbf{d}) - f(\mathbf{d}))^t S^- (f_n(\mathbf{d}) - f(\mathbf{d})) \tag{4.3}$$

where S^- is the (Moore-Penrose) generalized inverse of the sample covariance estimator computed on the samples. Note that the parameter ϵ needs to be set for M to be defined. The asymptotic distribution of NM can be shown to be chi-squared [3].

Note that we have defined the statistic in terms of the "densities", $f(\mathbf{d})$ and $f_n(\mathbf{d})$, but that we could equally well define the M statistic directly in terms of the cdfs computed at the same values $d_1, \ldots, d_m$. The two forms with, of course, appropriate definitional changes in the covariance matrix, yield identical results.

The cutoffs at which the interpoint distance distribution is evaluated can also be chosen so that between each two subsequent cutoffs one has, say, 10% of the probability mass. Such a choice can be expected to be more robust at the extremes.

Figure 4.2 shows the QQ plots for the null distribution of the M statistic for varying numbers of points versus a chi-squared random variable. The QQ plots are based upon 1000 realizations of the M statistic with points randomly distributed in the unit circle in the plane. The degrees of freedom used were the number of bins minus one. When N is 10, one of the bins is always empty, due to the smaller number of distances in this case, so we used eight degrees of freedom for the chi-squared distribution. These show that, except for the extreme right tails, the asymptotic results given above provide a good approximation to the distribution of M, even for small values of N.

Note that if one uses equal probability histogram, then the ecdf is such that the area of the empirical histogram of the frequency distribution of the interpoint distances converges to 0.1 within each bin. However, it is *not* true in general that the centered and scaled histogram converges in distribution to a multinomial distribution. The two discrete examples that follow illustrate this, and show that in situations of symmetry the rate of convergence of $F_n(d)$ to $F_D(d)$ may not be the one that one would expect.

4.3 Cluster Detection

To study the power properties of various detection statistics we designed a simple study with points generated at random on a unit circle in the plane. The number of points generated were determined by a Poisson distribution with mean 25. This represents the steady state. Outbreaks were then superimposed on this null distribution in the form of clusters of various sizes and locations. The power calculations are based on 1,000 repetitions.

We first looked at detection based solely on observing clusters based on the interpoint distances. One can compare the distributions using the classical Kolmogorov-Smirnov test

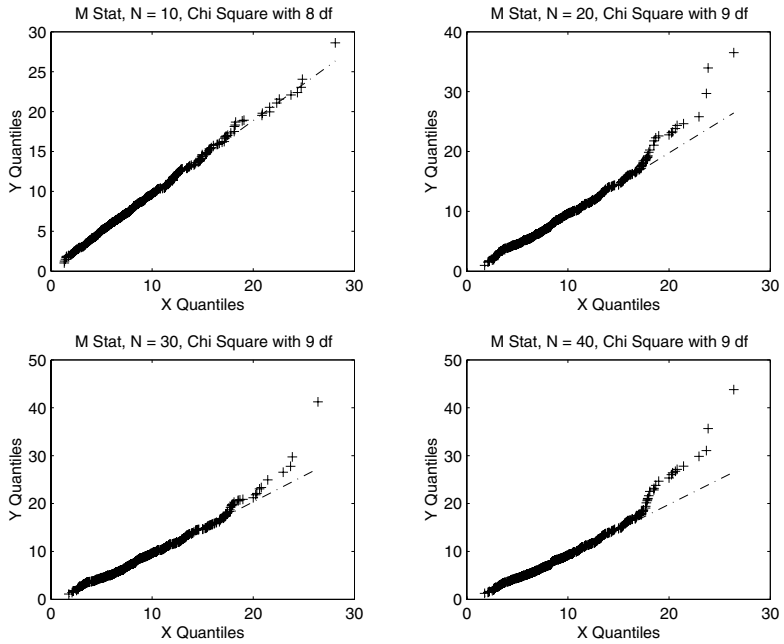

Figure 4.2. *QQ Plots of the null distribution of the M statistic versus the chi-squared distribution when points are randomly distributed in the unit circle, based on* 1, 000 *realizations of the statistic.*

statistic or the Wilcoxon test statistic (see, for example, [5]). The Kolmogorov-Smirnov test considers the largest difference between the two sample cdfs. The Wilcoxon test is rank-based and considers ranking within a combined sample of the two populations. The statistic is given by

$$W = \sum_{j=1}^{N} S_j$$

where the S_j are the ranks for one of the samples and N is the total number of distances in the combined sample. Because of the dependencies in the distances between patients, we cannot rely on published tables to determine the p-values of the tests. So we turned to simulations to derive empirical null-distributions of the statistics (see [4]). In our example we used 1000 samples to generate this distribution and determine a cutoff value corresponding to a Type I error rate of 0.05.

The tests that utilized the Wilcoxon and Kolmogorov-Smirnov statistics give very similar results for the problem at hand. Both tests are quite sensitive to the location of the cluster. In our example, we see that the power for detecting a cluster declines as the cluster moves farther from the origin. Figure 4.3 illustrates the results for the Wilcoxon test. These tests are sensitive to cluster location because of the impact the location has on the sample cdf. When a cluster is placed at the center of the circle the number of very small distances increases, but as the cluster moves to a more extreme location on the circle, the number of larger distances also increases. As a result, the alternative cdf becomes bimodal, and as neither of these tests has large power against such a bimodal distribution, they fail to detect such clustering.

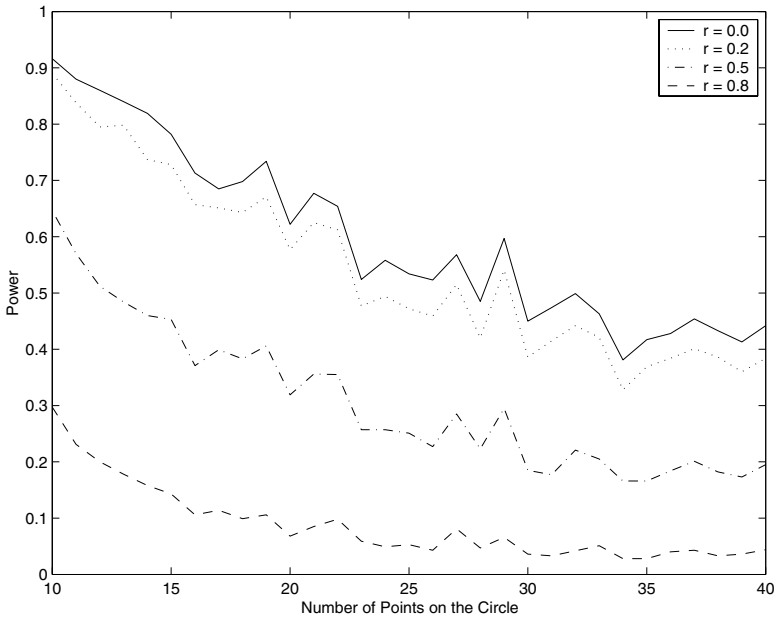

Figure 4.3. *Powers for the Wilcoxon Test with varying locations of the cluster. Cluster size* $= 5$, *radius* $= 0.001$, *location* $= 0.8$.

To address this lack of power, we choose to discretize the cdfs by placing the sorted interpoint distances into ten equiprobable bins. These bins are determined by the empirical distribution of distances of the null distribution. Subsequently, we can compare the numbers falling in each bin. To this end consider the M statistic as defined in equation 4.3, above. When using the M statistic in this example, we see an increase in the power to detect a cluster, regardless of its location on the circle (Figure 4.4), in contrast to the other two test statistics previously considered.

The Wilcoxon statistic seems to fail in the case where the cluster is at an extremity on the circle because the sum of the ranks for the two groups appear similar. This is because the statistic for the group with the cluster will tend to be composed of the ranks of the lower and higher points in the combined sample while the other group will be mainly composed of the intermediate ranks (the bi-modality effect). Therefore the overall sums of the ranks will be very similar.

The Kolmogorov-Smirnov statistic only considers the maximal difference between the cdfs. This summary ignores substantial information about the overall behavior of the groups in relation to one another. When a cluster is added to an extreme location on the circle, the difference between the two cdfs is divided between differences that occur in the smaller distances and the larger distances. Therefore the maximal difference does not appear to be so extreme and only captures one of the aberrations created by this case. The M statistic does not suffer from either of these shortcomings as illustrated above.

In what follows we consider some of the properties of the M statistic in order to better understand its distributional properties so as to enhance our capabilities for inference.

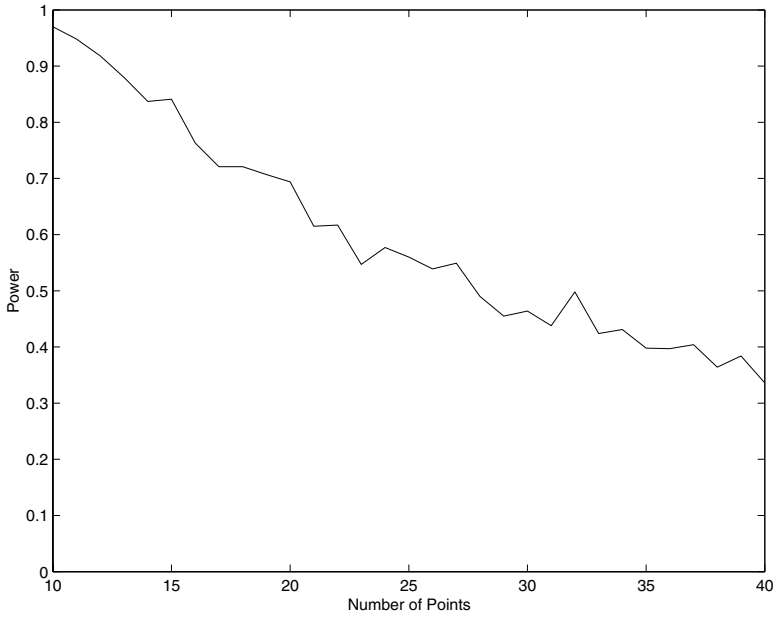

Figure 4.4. *Power of the M statistic to detect a cluster. Cluster size* $= 5$, *radius* $= 0.001$, *and various locations indicated by r in the legend; r varies from* 0.0 *to* 0.8.

4.3.1 Bivariate Power Calculations

We have seen the effectiveness of the M statistic at detecting clusters generated on a unit circle. As described previously, our aim is to be able to combine this test with a test of a Poisson process in order to more accurately and rapidly detect aberrations in a system that considers the number of events occurring as well as the spatial clustering of those events.

Clearly, the number of cases of a disease are the determinant of whether an outbreak has occurred or not. When there is value in early detection of an outbreak, other useful information can be incorporated to facilitate more rapid detection. So the task at hand is to *simultaneously* consider the number of cases *and* the location of these cases. Combining these two streams of information should improve our ability to detect outbreaks.

A very simple first step analysis is to consider that N, the number of cases, follows a Poisson distribution with parameter λ. Then one can superimpose an outbreak on this.

We know that asymptotically, conditional on n, as $n \to \infty$, that nM is distributed as a chi-squared variable [3], whose parameters are independent of n. This motivates us to think of n and nM as independent. This, in turn, suggests a bivariate rejection region such as shown in Figure 4.5 where the two pieces of information are combined so that the overall Type I error rate is 0.05. This compromise region protects against either too many cases or too much clustering individually, or in concert.

We continue with the example previously described, but now additionally assume that the number of sample points follow a Poisson distribution with a mean of 25. Table 4.1 gives a summary of the bivariate powers when the poisson and M statistic are combined, and contrasts these to the situation when we only use the M statistic. One can see that this method of combining the two statistics is not optimal. Indeed, an optimal boundary may depend on where the cluster is placed. But as a compromise boundary we see that the power

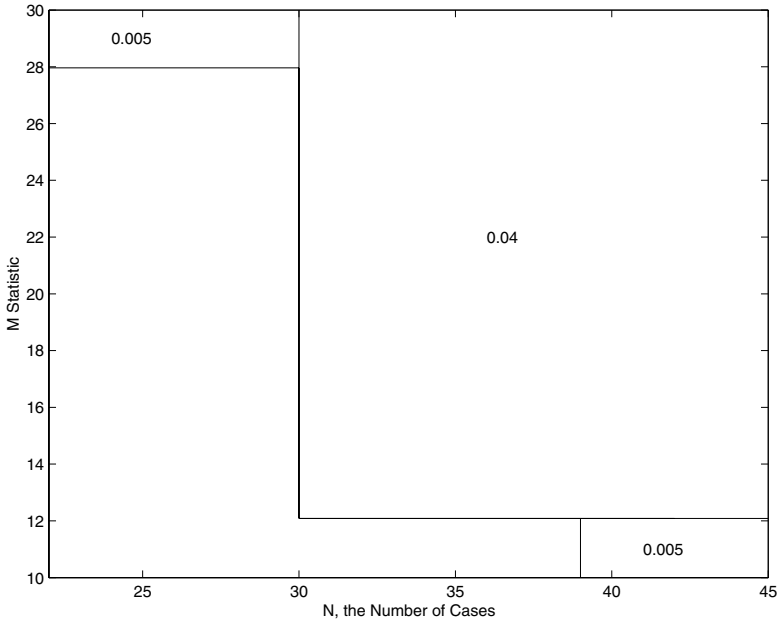

Figure 4.5. *Critical region for the bivariate test, a combination of a region where n is too large, where M is too large, and where both are too large, simultaneously. The probabilities of lying in each region under the null hypothesis are shown.*

is not too influenced by the location of the cluster.

4.3.2 Use of Nearest Neighbors Distances

Traditional methods for using distances to detect clustering have involved the nearest neighbor (see [6], for example). The M statistic above uses all distances, so the question naturally arises as to whether one loses any information by only looking at the closest neighbor. We report on a small study that looked at the power of detecting a cluster based on looking at the k-nearest neighbors. This tends to simplify the problem and minimize the amount of dependencies in the data, but it suffers from other weaknesses. Table 4.2 illustrates the results from doing this with our usual problem. Here the cluster is of size five, located at radius 0.5 and of radius 0.001. We see that there is no advantage to using near neighbor data, in fact it might compromise power, depending on the size of the cluster. Since we would not anticipate knowing the size of the cluster at the time of testing, this method does seem to be optimal.

Therefore, for this simple case, we have a method of detecting spatial and quantitative aberrations in a system. By utilizing the M Statistic to detect clustering and combining it with information gained from the Poisson distribution, one can powerfully detect an outbreak. It is possible to extend this technique for use in more complicated situations, such as in populations where the cases are not expected to occur uniformly over the area of consideration. One also might consider the case where multiple addresses are used, as will be described later or the case when exact addresses are not available.

Table 4.1. *Power to detect clusters at various locations in the unit circle when the clusters are superimposed on a Poisson (mean 25) number of points. The columns headed M statistic are the powers when considering the M statistic alone. The columns headed Joint are the powers when using the bivariate statistic.*

Cluster Location	Cluster size 5 Poisson power = 0.24		Cluster size 8 Poisson power = 0.45	
	M statistic	Joint	M statistic	Joint
0.0	0.7798	0.6156	0.9254	0.9107
0.2	0.5877	0.5680	0.8861	0.9497
0.5	0.4974	0.5602	0.8554	0.9071
0.8	0.4731	0.5021	0.8504	0.9054
alpha	0.05	0.0474	0.05	0.0474

Table 4.2. *Bivariate power calculations when using nearest neighbors.*

Number of near neighbors	Power
1	0.2054
2	0.2078
3	0.5031
4	0.4774
5	0.6645
6	0.5834
7	0.6279
8	0.6226
9	0.5387
All	0.5602

4.3.3 Discretization of Addresses

Often the only data that a medical institution will release are zipcodes of the patient. The following simulates such a situation by discretizing the points generated on a uniform circle into sectors. Two methods for doing this are described below.

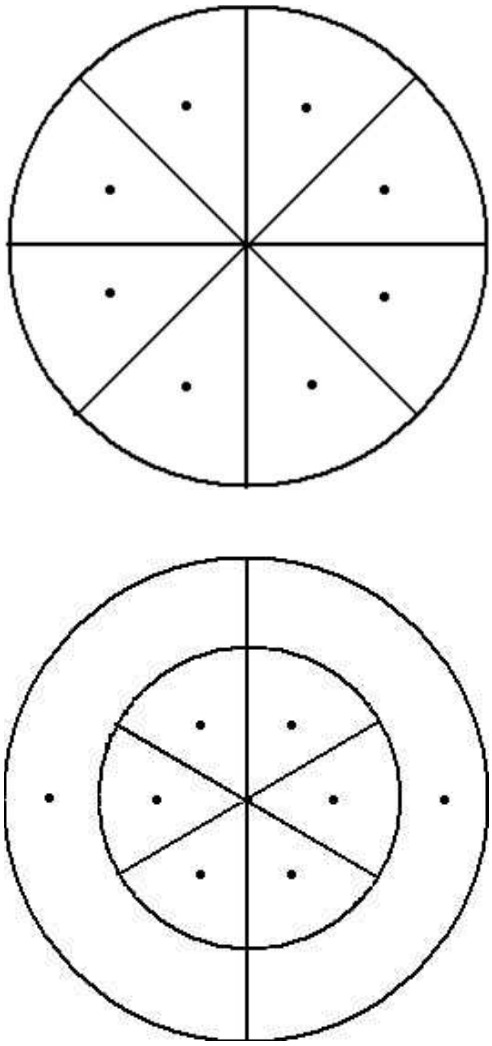

Figure 4.6. *The two graphs show the regions broken down into separate sectors. Anyone within a particular sector is reported as coming from the point displayed within that sector.*

Single Circle

The addresses were discretized into eight sectors in this simulation study. The sectors created are pie-shaped and all cases falling in a given sector are reported as coming from a point that is central to the sector. All of these points lie a distance of 0.5 from the center and are equally spaced (the left graph in Figure 4.6). The M statistic and the Poisson, as described in the previous section, were used to calculate the powers that are shown in Table 4.3. We considered three different locations of clusters: First we placed a cluster at the origin, then we centered a cluster on a boundary between two sectors, and finally we placed the cluster entirely in one sector. The cluster at the center had the potential of being split amongst any of the sectors. The one on a boundary had its impact split between two sectors. Thus the

Table 4.3. *Bivariate powers for the simple discrete case.*

Cluster Center	Cluster size 5	Cluster size 8
Poisson power =	0.24	0.45
0.0	0.2721	0.4880
On boundary	0.2449	0.3076
Contained in One Sector	0.2796	0.6842
alpha	0.0435	0.0435

Table 4.4. *Bivariate powers for the two-step discrete case.*

Cluster Center	Cluster Size 5	Cluster Size 8
0.0	0.3101	0.5188
On boundary		
Inner Circle	0.3202	0.6582
Outer Circle	0.4140	0.6656
In one Sector		
Inner Circle	0.3101	0.6761
Outer Circle	0.2493	0.5482
alpha	0.0483	0.0483

third placement was expected to have the most noticeable impact, as indeed it did.

There is a clear decline in power due to the discretization of the addresses. This may be partly due to the regular design we used that also greatly diminished the distinct number of distances. It is unusual that the powers for the cluster placed at the origin, where the points in the cluster could potentially fall into any one of the eight sectors, would be equivalent to the case when the cluster is entirely contained within one sector.

Two-Step Discrete

One might also consider a case similar to an urban setting, where the population density is largest at the center and diminishes as we move away from the center. One way to portray this is as follows. The circle is divided into two concentric circles, with the radius of the inner circle being 0.7. Six sectors are created in the inner circle while two are used on the outer circle (see the right figure in Figure 4.6). We considered five different cluster locations as described in Table 4.4.

We were drawn to considering what happens when an individual's address is coarsely

recorded, such as only getting the zip code; possibly for privacy reasons. One cannot draw too many general conclusions because of the very special design we used, with possibly too many regularities. These regularities restrict the number of distinct values the distances can take and thus the distribution of distances is now quite discrete with some large steps. But it is instructive to see what a large impact discretizing the data can have. More work needs to be done in this area.

4.4 The use of multiple distances

Conceptually, the consideration of addresses is appealing because it brings more information to bear on the problem. The simulations reported above are useful if the formation of the clusters has some connection to the point at which individuals were infected. Practically, we can use the individuals' home addresses to serve as proxies for the points of infection, admitting, of course, that this may not be the best proxy. Thus if more than one address is available per individual, such as an individual's home address and or work address, then it may prove beneficial to use these multiple addresses. Building on the statistical framework that we have already developed, we now consider this problem of incorporating multiple addresses and other data into the distribution of distances.

More formally, suppose for each individual we record a set of data, taken from a sample space $\mathcal{D}$. We have seen examples where $\mathcal{D}$ was discrete (section 2), and where $\mathcal{D} = \mathbb{R}^2$ (the address of the individual, as in sections 2, 3). We now allow $\mathcal{D}$ to be a higher-dimensional space, perhaps the product of several copies of $\mathbb{R}^2$ corresponding to multiple addresses, or a combination of discrete and continuous data. We will call any map $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ a *distance function*, understanding that this is an abuse of terminology (indeed, we do not require that our "distance function" take only positive values, satisfy the triangle inequality, or place any other restrictions on d).

In actual practice, the space $\mathcal{D}$ will be determined by the available data. The distance function d should make some attempt to model proximity in the context of the data and the problem at hand. We give some examples below.

1. Let $\mathcal{D} = \mathbb{R}^2 \times \mathbb{R}^2$. Each component indicates a separate recorded address. Possible distance functions between two data points include: taking the minimum of the (Euclidean) distance between first address and second address; taking the minimum of all possible address comparisons; or a weighted average of distance between first address and second address.

2. Let $\mathcal{D} = \mathbb{R}^2 \times \mathbb{R}$. The first component indicates a recorded address, and the second component a time of event. An appropriately chosen distance function can measure various spatial-temporal clustering effects (see [8], [9] and [7] and references therein, for example). As one example, given two data points (x_1, y_1, t_1) and (x_2, y_2, t_2), set $\Delta x = |x_1 - x_2|$, $\Delta y = |y_1 - y_2|$, and $d = |t_1 - t_2| \cdot \sqrt{(\Delta x)^2 + (\Delta y)^2}$. This measures proximity as a product of spatial and temporal distance.

3. Let $\mathcal{D} = \mathbb{R}^2 \times S$, where $S = \{1, 2, \ldots, n\}$ is a finite or discrete countable set. Given two data points (x_1, y_1, s_1) and (x_2, y_2, s_2), set $d = \sqrt{(\Delta x)^2 + (\Delta y)^2} + \alpha_1 \cdot \delta(s_1, s_2)$. Here, the delta function $\delta(s_1, s_2)$ takes the value 1 if $s_1 = s_2$, and 0 otherwise. The parameter α_1 is a real constant.

To illustrate the importance of using all of the available information, we ran a series of power calculations based on the last example above. Our simulated data consists of

Table 4.5. *Power to detect a cluster of size five in the first quadrant when the others in the sample are uniformly distributed in the unit circle. The contrast is between using solely the home address, only the school address, or both addresses.*

Sample size	10	15	20	25	30	35	40
Address only	0.40	0.21	0.20	0.14	0.13	0.12	0.10
District only	0.51	0.31	0.23	0.19	0.19	0.17	0.14
Address and district	0.71	0.41	0.34	0.23	0.19	0.19	0.18

individuals with address coordinates generated uniformly from the unit circle. Each quadrant of the circle corresponded to a hypothetical school district, and so each individual was assigned to a school according to the quadrant containing the individual's address. Now $\mathcal{D} = \mathbb{R}^2 \times \{1, 2, 3, 4\}$.

We then simulated disease outbreak in the following way. We generated a random sample of size between 10 and 40. We then reassigned five of these individuals to a broad region in the first quadrant (uniformly distributed throughout $R = \{(\rho, \theta) \,|\, 0.1 \leq \rho \leq 0.7, 0 \leq \theta \leq \frac{\pi}{2}\}$). The intention was to simulate an undetected attack at a school; because the residential addresses of affected children would be widely dispersed, usual methods of cluster detection may lack sensitivity to this pattern of disease.

The interpoint distances were calculated, first using the Euclidean distance only and ignoring the school district. We then used a chi-squared test (3 degrees of freedom) on the expected and actual counts of individuals in each school district, ignoring the address. Finally, we used both available components of the data, computing the interpoint distances with a distance function d as described in example 3 above. The univariate test statistic used on the interpoint distance distributions was the M statistic.

The power to detect these events of 5 individuals in one school district was greatest when both components of data were utilized. Gain in power was on the order of $25 - 75\%$. Table 4.5 below summarizes the power results:

When working with the distribution of interpoint distances, the outcome of any test statistic will depend on the choice of the distance function d. For statistics such as the M statistic (section 3) which rely on a binning procedure based on quantiles, there are equivalence classes of distance functions that leave the statistic invariant. Call two distance functions $d_1, d_2 : \mathcal{D} \to \mathbb{R}$ *monotonically equivalent* if $d_1 = \phi \circ d_2$ for some monotonic (increasing or decreasing) function $\phi : \mathbb{R} \to \mathbb{R}$. Then we have:

Fact: The M statistic is invariant across each equivalence class of monotonically equivalent distance functions.

Indeed, suppose $X_1 \ldots X_n$ are observed interpoint distances distributed according to a cdf F_X. Write $q_i = F^{-1}(i/100)$ for the quantiles of F_X. Suppose ϕ is a monotonically increasing (decreasing) function from $\mathbb{R}$ to $\mathbb{R}$. Then for any particular observation X_i that lies between q_j and q_{j+1}, monotonicity guarantees that $\phi(q_j) \leq \phi(X_i) \leq \phi(q_{j+1})$ (or reverse the inequalities for ϕ decreasing). Then bin the distances X_i into deciles of F_X as described above. This yields the same bin counts as when we bin the distances $\phi(X_i)$ and use the deciles of $F_{\phi(X)}$. Hence the value of M remains unchanged after transformation $X \mapsto \phi(X)$.

Example: Let $\mathcal{D} = \mathbb{R}^2$, and $d : \mathcal{D} \to \mathbb{R}$ the ordinary Euclidean distance. Let $\phi_1 = \frac{1}{x} : \mathbb{R} \to \mathbb{R}$ and $\phi_2 = \frac{\exp(1/x)}{1+\exp(1/x)} : \mathbb{R} \to [0, 1]$. Then for a fixed set of data, the M statistic will take the same value whether computed using distance function d, $d_1 = \phi_1 \circ d$, or $d_2 = \phi_2 \circ d$.

We find that using the transformation suggested by d_2 above is convenient for maintaining a normalized measure of proximity, taking values in [0, 1]. Note that we have inverted the scale of values, so that proximity close to 1 indicates two individuals that are very similar, while a value close to 0 indicates to individuals that are dissimilar. Depending on the model, we may gain some interpretability using this similarity measure. One possible interpretation might be $d_2 =$ the probability that two individuals became infected from the same source of disease. In practice we would first determine the available data and formulate an appropriate method of measuring proximity that captures the essence of the problem at hand. Call this distance function d, and applying ϕ_2 to d gives a proximity or similarity measure that takes values between 0 and 1.

4.5 Conclusions

We have attempted to show how to use the distribution of the interpoint distance between two randomly selected observations as a summary of a spatial distribution. It can be used in biosurveillance if it provides a sufficiently stable constancy in order to define normal behavior against which deviations can be spotted. That has been our experience. We have observed the distributions of distances between patients arriving at the Emergency Department with flu-like symptoms at a children's hospital in Boston, and they display a remarkable constancy over time [19].

The distribution can be estimated from the dependent distances between observations in a random sample generated from an underlying spatial distribution. It can then serve as the null distribution against which the deviations of future samples can be compared. The M statistic is an example of a derived statistic that is specifically designed to detect these deviations.

We show that the combination of the M statistic (or of another statistic based on the interpoint distance distribution) with a statistical test for the presence of an excessively high number of cases of some disease in a given time frame (in a day or week, for example) allows for an increase in the power to detect outbreaks caused by naturally occurring or deliberately released agents.

Given the many possible routes of infection in the event of an outbreak, the possibility of extending the concept of interpoint distance to a dissimilarity measure—in particular, through the use of the multiple addresses usually associated with each individual—allows for a straightforward generalization of the methods described here. Further extensions also include the fitting of models for the interpoint distance, also from dependent quantities obtained from all pairs of observations.

The development of surveillance systems that collect real-time information on health-related events (such as flu-like symptoms in pediatric emergency room admissions) and that use detection methods such as the ones introduced here should become a priority both for its public health and for its national security implications.

Acknowledgments

This research was funded in part by National Institutes of Health grants RO1AI28076 and T32AI07358.

Bibliography

[1] Reis B.Y., Pagano M., and Mandl,K.D. Using Temporal Context to Improve Biosurveillance, *Proc. Nat. Acad. Science* 2003;**100(4)**:1961-1965.

[2] Meselson M, Guillemin J, Hugh-Jones M, Langmuir A, Popova I, Shelokov A, Yampolskaya O,et al. The Sverdlovsk anthrax outbreak of 1979. *Science* 1994;**266**:1202-7.

[3] Bonetti, M. and Pagano, M. The interpoint distance distribution as a descriptor of point patterns: An application to cluster detection. *Submitted*, 2003.

[4] Dwass M. Modified Randomization Tests for Nonparametric Hypotheses, *Annals of Mathematical Statistics*, **28**,1957, pp. 181-187.

[5] Hollander and Wolfe, *Nonparametric Statistical Methods*, John Wiley & Sons., 1999.

[6] Cressie, N.A.C. *Statistics for spatial data* Wiley-Interscience, 1991.

[7] Jacquez, G. M. A *k* nearest neighbour test for space-time interaction, *Statistics in Medicine* **15**,1996, 1935–1949.

[8] Knox, G. The detection of space-time interactions, *Applied Statistics* **13**, 1964,25–29.

[9] Mantel, N. The detection of disease clustering and a generalized regression approach, *Cancer Research* **27**, 1967, 209–220.

[10] M.S. Bartlett. The spectral analysis of two-dimensional point processes. *Biometrika*, 51:299–311, 1964.

[11] E. Borel. *Traité du Calcul des Probabilités et de ses Applications, I*. Paris: Gauthier-Villars, 1925.

[12] Víctor de la Peña and Evarist Giné. *Decoupling*. Springer-Verlag, 1999.

[13] M. Dwass. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28:181–187, 1957.

[14] J.E. Kelsall and P.J. Diggle. Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine*, 14:2335–2342, 1995.

[15] Whitney K. Newey and Daniel McFadden. *Large sample estimation and hypothesis testing*, pages 2113–2241. Elsevier/North-Holland, 1994.

[16] B. W. Silverman, Limit theorems for dissociated random variables. *Advances in Applied Probability*, 8:806–819, 1976.

[17] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[18] B. S. Weir, *Genetic Data Analysis II*. Sinauer Associates, Inc., 1996.

[19] K. L. Olson, M. Bonetti, M. Pagano, and M. D. Mandl, Syndromic Surveillance: A Population-Adjusted, Stable Geospatial Baseline for Outbreak Detection, *Proceedings of the AMIA Fall Symposium* 2003 [submitted]