8

Distance-Based Methods for Spatial and Spatio-temporal Surveillance

Laura Forsberg, Marco Bonetti, Caroline Jeffery, Al Ozonoff and Marcello Pagano

8.1 INTRODUCTION

The emergence of new infectious diseases and the threat of biological attacks have lead to a growing interest in methods of surveillance, including the accompanying statistical methods, for the early detection of an outbreak. Statistically, what we would like to do is detect the time point at which there is an increase in the number of infected individuals, an increase that may also be accompanied by a change in the spatial distribution of these patients, either, or both, of which might indicate an outbreak of some sort – a disturbance of normalcy. The time element is critical in that a less than timely detection would make the methods essentially useless.

The timeliness is an extra consideration that possibly distinguishes the newer surveillance methods from those in the older literature. The older ones are often related to such issues as the detection of cancer clusters (see, for example, Alexander and Boyle, 1996), and sometimes use data that was collected over a period of years prior to analysis which, parenthetically, makes the existence of a cluster questionable. This is not meant as a criticism of the classical methods as the time element is inherent in those methods, too.

When considering spatial methods for cluster detection, no method seems to be uniformly better than all others, so it is beneficial to review a number of these methods. Several reviews of statistical methods for the detection of spatial anomalies have been written (see, for example, Kulldorff, 1998; Elliott *et al.*, 2000;

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman
 ISBN 0-470-09248-3 (HB)

Lawson, 2001; Brookmeier and Stroup, 2004, Chapter 7). Most of the statistical methods that have been described for the detection of spatial anomalies can be grouped into two general categories: quadrat methods and distance methods. Quadrat methods divide the geographical region into smaller areas termed quadrats and compare the incidence of events within the quadrat to the incidence in the remaining study region. The spatial scan statistic is perhaps the most widely known and used of these methods (Kulldorff, 1997). Distance-based methods, on the other hand, consider some measure of distance between events. Usually Euclidean distance is used as the measure of distance between individuals, but typically any measure of dissimilarity or similarity between events can be utilized. We focus on these distance-based methods in this chapter, and discuss two methods of more modern interest: the maximized excess events test (MEET) and the M statistic, with particular emphasis given to the latter method. We present the motivation for using distance-based methods in Section 8.2. In Section 8.3, we give a review of the MEET statistic and the M statistic, and their utility in public health surveillance. Section 8.4 introduces a data example to illustrate the implementation of the MEET and M statistic to detect spatial clusters of disease. Our focus of attention is a bivariate statistic, which simultaneously monitors case volume and the spatial distribution of the cases. This bivariate statistic is introduced to improve the power to detect suspicious patterns in the data stream. In Section 8.5, we describe and illustrate a method for determining the location of a cluster, or other spatial aberrations, once the M statistic has indicated that such an anomaly exists.

8.2 MOTIVATION

Distance-based methods consider the distribution of the pairwise interpoint distances between all the individuals in the study region. Under the null hypothesis this distribution remains stable. As time progresses, we need to be on the alert for a disturbance in the distribution. This alternative distribution should be sensitive to the detection of disturbances in how individuals are located, especially if individuals are clustered. These disturbances are those we would expect during an outbreak of a contagious disease or an outbreak resulting from one or more point source emissions of some bioterrorist agent. This places the problem in the classical hypothesis testing paradigm, and to pursue this thinking further, we seek methods that will have power against alternatives that reflect clustering of individuals. One obvious characterization of clustering is to consider pockets of individuals who consequently will have smaller average distances between themselves than they would in the null case. But this is not the only alternative one can envision; others may impact the second moment of the distribution of distances, for example.

One method considers a test of the mean of the interpoint distance distribution (Whittemore *et al.*, 1987). The statistic, usually called the δ statistic, is equal to a

weighted average of the observed distances, and thus tests for shifts in the mean of the interpoint distance distribution. Subsequent work has shown that this method is not very powerful at detecting clusters (Bonetti and Pagano, 2004a). The reason for the lack of power is that the mean is not an efficient summary of the null distribution, typically because the null distribution of distances is not normal. Furthermore, dependencies in the distances can often lead to complex deviations from the null distribution that may not necessarily lead to a shift in the overall mean. Figure 8.1 illustrates such a scenario arising from real data. Here the densities clearly differ from one another, but the mean does not lead to a powerful statistical test for detecting such a deviation.

Dealing with distances between individuals requires some thought since the usual statistical methods do not apply seamlessly. First, the distances themselves are not independently distributed. This would seem clear considering that for every n individuals there are $\binom{n}{2}$ distances. Thus considering the statistical properties of their joint distribution is not straightforward. Additionally, location data is often not reported precisely, but rather it is reported in a discretized manner. For instance, instead of individuals' home or work addresses we may only be told the census tract, postal code, or county in which they reside. Thus the distances can only assume values in a finite grid.

Additionally, the location of spatial aberrations in the study region will impact the shape that the alternative distribution will assume. For instance, a cluster

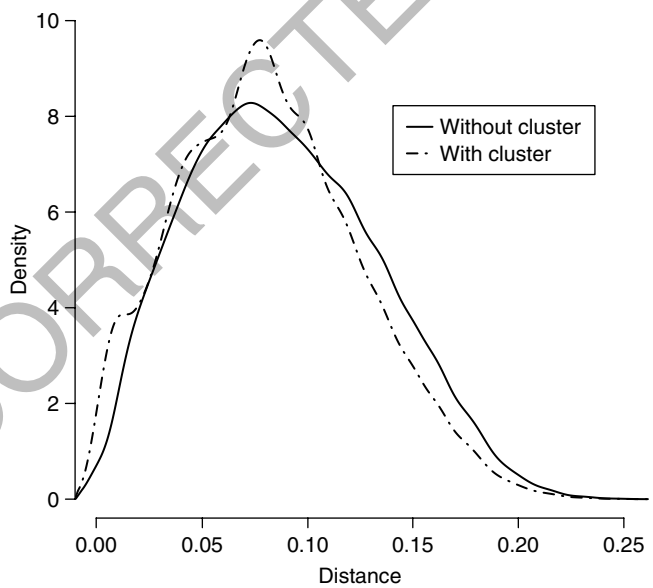

Figure 8.1 Distribution of the distances for a data set with no clusters ($\hat{\mu} = 0.090, \hat{\sigma} = 0.045$) versus a the same data set with clusters superimposed ($\hat{\mu} = 0.083, \hat{\sigma} = 0.044$).

placed in the study region will create a larger than expected number of small distances. However, the cluster will also create other abnormalities in the distribution, but these will depend upon where the cluster is placed, due to the addition of the distances between the cluster and other points in the region. This patterning increases the more clusters we have.

Several methods for analyzing distances have been proposed, although no one statistic seems to completely handle the complexities that distance data presents uniformly better than any others. K functions are one method that has been proposed (Ripley, 1976; Diggle and Chetwynd, 1991) for detecting spatial abnormalities, especially in the ecological literature (Dobbertin *et al.*, 2001; Couteron and Kokou, 1997). These functions enjoy nice mathematical properties, but can be cumbersome to implement for purposes of biosurveillance. Therefore we will direct our attention to two other methods, the MEET statistic and the M statistic, with particular emphasis on the latter.

8.3 DISTANCE-BASED STATISTICS FOR SURVEILLANCE

8.3.1 MEET Statistic

Tango (1995) describes a method of cluster detection that assumes that the data is aggregated into m regions according to some spatial boundaries, for instance by zip code or county. The statistic considers the difference between the observed rate of cases in each region and the expected rate, and then weights these differences by a measure of the distance between the regions. More explicitly, within the ith region, let y_i be the observed number of cases and e_i be the expected number of cases. Define the parameter λ such that any pair of cases that are farther than λ apart cannot be considered a cluster. Basically, λ can be thought of as some measure of the spatial extent of a cluster. Consider the vectors $\mathbf{r} = \{r_i\}$, where $r_i = y_i / \sum_{i=1}^{m} y_i$, and $\mathbf{p} = \{p_i\}$, where $p_i = e_i / \sum_{i=1}^{m} e_i$. Then the estimated events test statistic is given by

$$C_\lambda = (\mathbf{r} - \mathbf{p})^{\mathrm{T}} \mathbf{A}(\lambda)(\mathbf{r} - \mathbf{p}),$$

where $\mathbf{A}(\lambda) = \{a_{ij}(\lambda)\}$. One can consider several forms for the $a_{ij}(\lambda)$. Clearly, the choice of the form that $\mathbf{A}(\lambda)$ assumes will have an impact on the efficacy of this statistic. However, the magnitude of this effect and the sensitivity of the statistic to $\mathbf{A}(\lambda)$ have not been studied systematically. In practice the exponential threshold model has been used (Tango, 2000), such that $a_{ij}(\lambda)$ is defined as

$$a_{ij}(\lambda) = \exp\left\{-4\left(\frac{d_{ij}}{\lambda}\right)^2\right\},$$

where d_{ij} is the Euclidean distance between regions i and j. The problem with this method is that it requires specification of the parameter λ. Generally this is not known a priori, and several values of λ are tested, leading to multiple testing problems. In order to circumvent this problem, Tango (2000) developed the maximized excess events test (MEET). This statistic searches for the value of λ which gives the smallest p-value of the observed value of C_λ, denoted c_λ, as follows,

$$P = \min_{\lambda} \Pr\{C_\lambda > c_\lambda | H_0, \lambda\}.$$

This is implemented by allowing λ to assume discrete values near zero up to about half of the size of the study area and performing a line search over these values of λ. Monte Carlo simulation methods are used to obtain the null distribution of P.

8.3.2 The Interpoint Distribution Function and the *M* Statistic

The M statistic uses the interpoint distance distribution and its empirical cumulative distribution function (ecdf) to perform inference. Consider a spatial distribution $P(\mathbf{x})$ defined over a bounded region of the plane. Let the point distribution over the region be absolutely continuous, so that for two independent and identically distributed points $\mathbf{x}_1$ and $\mathbf{x}_2$ in the region, $P(\mathbf{x}_1 = \mathbf{x}_2) = 0$. For any such point distribution P, if one defines a nonnegative distance (or dissimilarity) function d, then the random variable $D = d(\mathbf{x}_1, \mathbf{x}_2)$ has some distribution $P_D(d)$. We call D the interpoint distance between two independent points. The cdf $F(\cdot)$ of D is $F(d) = EI(d(\mathbf{x}_1, \mathbf{x}_2) \leq d)$, where $I(\cdot)$ is the indicator function and E denotes expectation with respect to the $P \times P$ distribution.

Extending the usual definition of an ecdf for random samples, one can define the ecdf of the interpoint distances associated with a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as

$$F_n(d) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} I(d(\mathbf{x}_i, \mathbf{x}_j) \leq d).$$

The quantity $\sqrt{n}(F_n(d) - F(d))$, considered as a stochastic process indexed by d, converges weakly to a Gaussian process (Silverman, 1976; Bonetti and Pagano, 2004a). Because of the very definition of a Gaussian process, this general result implies that for a fixed value d the cdf $F_n(d)$ has $\sqrt{n}$-convergence to $F(d)$.

More generally, consider the empirical cdf $F_n(\mathbf{q}) = (F_n(q_1), \ldots, F_n(q_k))$ computed at a finite number k of fixed values $\mathbf{q} = (q_1, \ldots, q_k)$. The cutoff

points q_j are typically chosen to be the $(j/k)100\,\%$ percentiles of the distribution of D. If the range of D is unbounded, we set $q_k = \infty$. Then, the weak convergence implies that the joint asymptotic distribution of the centered ecdf $\sqrt{n}(F_n(\mathbf{q}) - F(\mathbf{q})) = \sqrt{n}(F_n(q_1) - F(q_1), \ldots, F_n(q_k) - F(q_k))$ is asymptotically multivariate normal with covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{a,b}\}$, with

$$\begin{aligned}\sigma_{a,b} \;=\; & E[I(d(\mathbf{x}_1, \mathbf{x}_2) \le q_a, d(\mathbf{x}_1, \mathbf{x}_3) \le q_b)] \\ & - EI(d(\mathbf{x}_1, \mathbf{x}_2) \le q_a) EI(d(\mathbf{x}_1, \mathbf{x}_3) \le q_b).\end{aligned}$$

A number of standard test statistics can be used to evaluate the distance between $F_n(\cdot)$ and $F(\cdot)$ for hypothesis testing, but the lack of independence among observed distances between individuals precludes the use of standard statistics without using appropriate modifications.

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

$$\tilde{M}(F_n(\mathbf{q}), F(\mathbf{q})) = (F_n(\mathbf{q}) - F(\mathbf{q}))^{\mathrm{T}} \boldsymbol{\Sigma}^{-} (F_n(\mathbf{q}) - F(\mathbf{q})),$$

a Mahalanobis-like statistic, where $\boldsymbol{\Sigma}^{-}$ is a generalized inverse (see Rao and Mitra, 1971) of the covariance matrix of the vector $F_n(\mathbf{q})$. For definiteness we use the Moore–Penrose generalized inverse. In applications we typically use an estimator of $\tilde{M}$: consider the quadratic form

$$M(F_n(\mathbf{q}), F(\mathbf{q})) = (F_n(\mathbf{q}) - F(\mathbf{q}))^{\mathrm{T}} \mathbf{S}^{-} (F_n(\mathbf{q}) - F(\mathbf{q})),$$

where $\mathbf{S}$ is the estimated covariance matrix, obtained by generating repeated samples of size n from an assumed null spatial distribution of the individuals over the region of interest. To calculate $\mathbf{S}$ we could also take repeated samples from historic data, if available. We note that the M statistic can also be computed when the data consists of counts recorded at a finite number of fixed locations (see Bonetti and Pagano, 2004a), with minor modifications. If these fixed locations are a result of a discretization of the individuals addresses, there is the possibility of a loss of power to detect deviations from the null geographic distribution.

An alternative definition of M can be given in terms not of the cumulative distribution function, but of its first differences at the subsequent bin counts along the distance axis. The ecdf and the cdf of D are therefore summarized by the observed proportions o_j and the expected probabilities $e_j = j/k$ within each of the bins, with $j = 1, \ldots, k$. The variance–covariance matrix in that case needs to be modified in the obvious manner, since the first differences are a linear combination of the values of the cumulative distribution functions.

As an alternative, a consistent estimator for the variance–covariance matrix $\boldsymbol{\Sigma}$ can also be constructed. The covariance matrix can be estimated consistently by the terms

$$\widehat{\sigma}_{a,b} = 4\left\{\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n} h(\mathbf{X}_i,\mathbf{X}_j,\mathbf{X}_k;q_a,q_b) - \left[\frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n} I(d(\mathbf{X}_i,\mathbf{X}_j)\le q_a)\right] \times \left[\frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n} I(d(\mathbf{X}_i,\mathbf{X}_j)\le q_b)\right]\right\},$$

where

$$h(\mathbf{X}_i,\mathbf{X}_j,\mathbf{X}_k;q_a,q_b) = 6^{-1}\sum_{\rho}[I(d(\mathbf{X}_{\rho_1},\mathbf{X}_{\rho_2})\le q_a, d(\mathbf{X}_{\rho_1},\mathbf{X}_{\rho_3})\le q_b)]$$

is the symmetrized kernel computed over the collection $\rho = \{(\rho_1,\rho_2,\rho_3)\}$ of the six permutations of the indices (i,j,k) (see Bonetti and Pagano, 2004b). In the calculation of this estimator, for efficiency the triple sum should be implemented as a single loop by making use of (fast) matrix multiplications for the inner sums.

8.3.2.1 Example

As an example, consider points uniformly distributed on the unit square $[0,1]\times[0,1]$. The distribution of the interpoint distance between two such points is as described in Bartlett (1964). The approximate quantiles at probabilities (0.2, 0.4, 0.6, 0.8, 1) from that distribution are (0.2912, 0.4435, 0.5891, 0.7573, 1.4142). Using these as cutoff values, consider the empirical estimator of that cdf $F_n(q_h)$ at the deciles $q_h, h = 1,\ldots,5$. Note that $q_5 = 2^{1/2}$ is the largest possible interpoint distance on the unit square, and that the cumulative distribution function is always equal to one for that value, so that consideration of $F_n(d_h)$ at $d_h, h = 1\ldots,4$ suffices.

Table 8.1 shows the asymptotic variance–covariance matrix ($\boldsymbol{\Sigma}^*$) of $n^{1/2}(F_n(d_1)-F(d_1),\ldots,F_n(d_4)-F(d_4))$, as estimated from 3000 samples of size 5000.

We then considered four sample sizes n = 100, 250, 500, and 1000. For each sample size we computed the estimator of the variance–covariance matrix $\boldsymbol{\Sigma}$ one hundred times, as described above. On the left-hand side of Table 8.2 we report, for each sample size and for each element of the matrix, the relative bias of the variance–covariance matrix estimator, computed as the difference between the average of the 100 matrices and $\boldsymbol{\Sigma}^*$, divided by $\boldsymbol{\Sigma}^*$. On the right-hand side of Table 8.2 we report, also for each sample size and for each element of the matrix, the coefficient of variation relative to $\boldsymbol{\Sigma}^*$, that is, the ratio between the standard deviation of each term as computed from the 100 matrices and $\boldsymbol{\Sigma}^*$.

The variance–covariance matrix estimator appears to be centered satisfactorily at the true (as estimated by $\boldsymbol{\Sigma}^*$) variance–covariance matrix of the ecdf of

Table 8.1 Estimated variance–covariance matrix $\boldsymbol{\Sigma}$ of $\sqrt{n}$ times the interpoint distance ecdf. The matrix is based on 3000 samples of size 5000.

	d_1	d_2	d_3	d_4
d_1	0.011	0.022	0.029	0.027
d_2	0.022	0.051	0.068	0.058
d_3	0.029	0.068	0.092	0.077
d_4	0.027	0.058	0.077	0.060

Table 8.2 Relative bias and coefficient of variation (relative to $\boldsymbol{\Sigma}^*$ in Table 8.1) of the estimator of $\boldsymbol{\Sigma}$.

Relative bias

$n = 100$

	d_1	d_2	d_3	d_4
d_1	−0.06	−0.05	−0.04	−0.03
d_2	−0.05	−0.05	−0.04	−0.02
d_3	−0.04	−0.04	−0.03	−0.02
d_4	−0.03	−0.02	−0.02	0.00

$n = 250$

	d_1	d_2	d_3	d_4
d_1	0.02	0.02	0.02	0.01
d_2	0.02	0.01	0.01	0.01
d_3	0.02	0.01	0.00	0.00
d_4	0.01	0.01	0.00	0.00

$n = 500$

	d_1	d_2	d_3	d_4
d_1	0.05	0.05	0.04	0.04
d_2	0.05	0.03	0.02	0.02
d_3	0.04	0.02	0.02	0.02
d_4	0.04	0.02	0.02	0.01

$n = 1000$

	d_1	d_2	d_3	d_4
d_1	0.04	0.04	0.03	0.03
d_2	0.04	0.03	0.02	0.02
d_3	0.03	0.02	0.02	0.02
d_4	0.03	0.02	0.02	0.02

Coefficient of variation

$n = 100$

	d_1	d_2	d_3	d_4
d_1	0.60	0.45	0.35	0.24
d_2	0.45	0.28	0.19	0.13
d_3	0.35	0.19	0.12	0.08
d_4	0.24	0.13	0.08	0.11

$n = 250$

	d_1	d_2	d_3	d_4
d_1	0.35	0.29	0.24	0.16
d_2	0.29	0.19	0.13	0.09
d_3	0.24	0.13	0.08	0.05
d_4	0.16	0.09	0.05	0.07

$n = 500$

	d_1	d_2	d_3	d_4
d_1	0.20	0.17	0.14	0.09
d_2	0.17	0.11	0.07	0.05
d_3	0.14	0.07	0.04	0.03
d_4	0.09	0.05	0.03	0.05

$n = 1000$

	d_1	d_2	d_3	d_4
d_1	0.13	0.11	0.09	0.06
d_2	0.11	0.07	0.05	0.03
d_3	0.09	0.05	0.03	0.02
d_4	0.06	0.03	0.02	0.03

the interpoint distance. The relative bias of the estimator is reassuringly small (less than or equal to 6 %) even for the smaller values of n. The variance of the estimator is such that the relative standard errors only fall below 20 % when the sample size n is at least equal to 500. Lastly, it should also be noted there there tends to be more bias and variability in the estimation of the variances and covariances that involve the cdf evaluated at small distances compared to larger distances.

8.4 SPATIO-TEMPORAL SURVEILLANCE: AN EXAMPLE

Although the focus of this chapter is on spatial methods, we may also consider the temporal aspect of a surveillance data stream, as well as methodology that integrates the spatial and temporal information for the purposes of surveillance. This integrated approach is often referred to as spatio-temporal surveillance. In this section we illustrate the spatial methods described above with a real data set, and then continue our example with this data set to illustrate the utility of temporal and spatio-temporal methodology. To simplify the exposition we only consider the day-to-day behavior of the system and ignore any memory from one day to the next. Clearly, a real system would have memory beyond a single day (see Reis *et al.*, 2003).

The data set that we use to illustrate these methods was collected by a large health provider in Massachusetts. As patients arrive for emergency care, their cases are geocoded (typically the residential or billing address of the patient), and this information is centralized electronically on a daily basis. In the interest of anonymity, in this exposition the spatial data provided has been aggregated by census tract, with jittering to further conceal the true locations of the individual patients involved. We consider a subset of these electronic data, consisting of upper respiratory infections (URIs) arriving at emergency and urgent care departments for this provider between the dates of January 1, 1996 and October 30, 1999, a stretch of 1399 days or nearly four years of daily data.

This data stream thus provides the temporal patterns of disease in the form of the number of cases arriving each day, as well as the spatial patterns of disease produced as the locations of patients change over time. For all further analysis, we have divided the data into three groups according to the day of the week: weekends and holidays, days after weekends and holidays, and the remaining days in the week. This was necessary because some of the locations provided were closed on weekends and holidays, leading to a stratification of case volume and spatial patterns on different days of the week.

Since there were no known bioterrorism attacks in Massachusetts during the period of study, for the purpose of evaluating methods, we chose to augment the real data with simulated clusters. To this end, we created three new data sets. For two, we chose six adjacent census tracts in close proximity and added one additional URI per day per tract, for a total of six additional cases per day. For

brevity, we call such a simulated signal a 'cluster'. The two data sets contain clusters centered around census tracts labeled 477 and 179 respectively, and we refer to the corresponding data sets accordingly. In a third round of simulations, we added both clusters of six cases, for a total of 12 additional cases (six each in the two separate locations; see Figures 8.2 and 8.6).

8.4.1 Temporal Component

Before consideration of the spatial and spatio-temporal surveillance of the Massachusetts data, we briefly describe an approach to the temporal surveillance of such data. Rather than describe the variety of methods available (see Chapter 2), we simply describe the modeling approach that we have taken with these particular data.

Let $N(t)$ denote the daily case volume of URIs across the entire study area, $1 \leq t \leq 1399$. The time series $N(t)$ shows several trends which make modeling challenging. Both the mean and variance of $N(t)$ have strong seasonal and day-of-week variation (see Figures 8.3 and 8.4). Closure of some locations on weekends and holidays further complicates modeling and analysis.

In order to construct a model for $N(t)$ we first used standard linear regression methods to fit a deterministic component for the mean expected case count. This is essentially the approach described in Brookmeier and Stroup (2004,

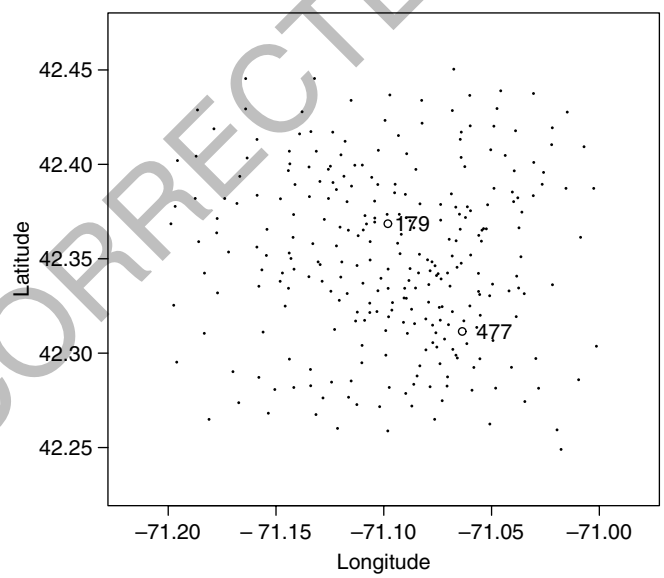

Figure 8.2 Locations of the census tracts with superimposed clusters, relative to the remaining census tracts in the Massachusetts data set.

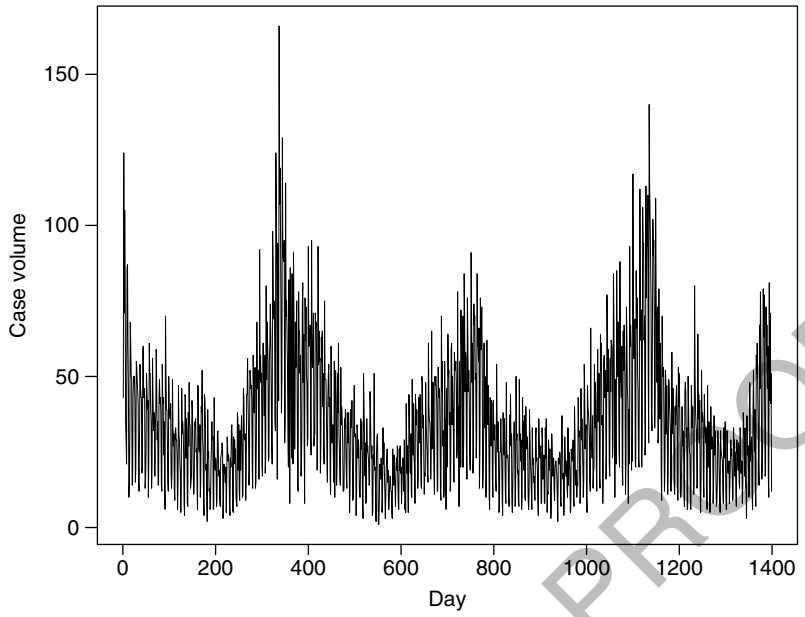

Figure 8.3 The time series $N(t)$ exhibits a seasonal pattern in addition to occasional sharp increases in the winter months.

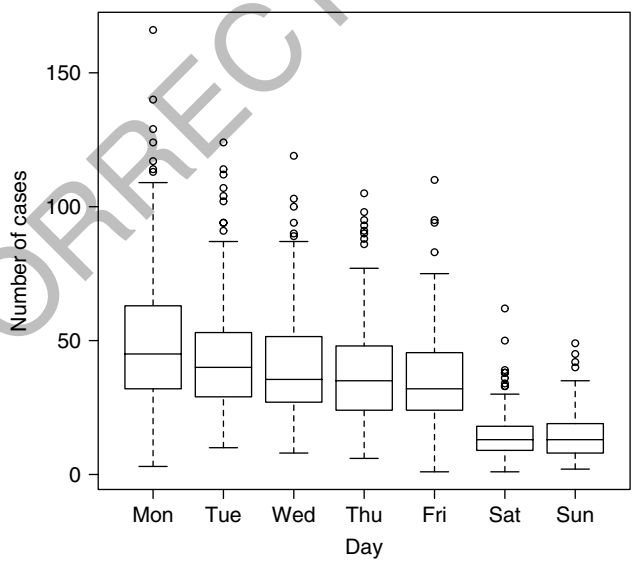

Figure 8.4 Number of cases by day of week.

pp. 203–231). The linear model included several harmonic terms for the characteristic seasonal effect on URIs, as well several indicator variables corresponding to identifiable day-of-week effects. An additional indicator for the months of December through February (the well-known 'flu season') was included to account for the frequent excess of cases in these months. The day-of-week variation is exhibited in both first and second moments, so after subtracting the fitted values from the observed data we divided by the daily standard error in order to standardize the residuals. Denote by $\eta(t)$ the time series constructed from each resulting data point; we can think of $\eta(t)$ as a standardized residual departure from the baseline mean.

The residuals $\eta(t)$ are characterized by a high degree of autocorrelation. Our goal is to model the residuals, resulting in a predicted value for $N(t)$ that can be compared to the observed value. Taking a simple approach, we used a first-order autoregression (AR(1)) to model the autocorrelation. After inclusion of the autoregressive terms the standard deviation of the residuals was reduced by nearly 10 % from 0.923 to 0.838. Thus the full model is:

$$\begin{aligned} N(t) \equiv\ & \alpha_0 + \alpha_1 \cos\left(\frac{2\pi}{365}\right) + \alpha_2 \sin\left(\frac{2\pi}{365}\right) + \alpha_3 I(\text{wkend}) \\ & + \alpha_4 I(\text{Monday}) + \alpha_5 \cos\left(\frac{2\pi}{30}\right) + \alpha_6 \sin\left(\frac{2\pi}{30}\right) \\ & + \alpha_7 I(\text{flu season}) + \text{interaction terms} + \in(t), \\ \eta(t) \equiv\ & \frac{\in(t)}{\sigma} = \beta\eta(t-1) + \xi(t). \end{aligned}$$

Thus we can view $N(t)$ as a test statistic for temporal surveillance, where we consider any observed N falling in a critical region to raise an alarm.

8.4.2 Bivariate Test Statistic

In order to fully utilize the available information, we consider using a bivariate test statistic, composed of the two statistics, the M statistic calculated daily and $N(t)$, described above. In an abuse of notation we refer to their daily product as NM, dropping the reference to time, t.

$N(t)$ allows us to calculate a residual value for the number of cases arriving, based on the time series prediction for that day, and the residuals are distributed approximately normally. Simultaneously, $\log(NM)$ can be used to evaluate the deviation of the spatial distribution of cases from normalcy. Theoretical research (Bonetti and Pagano, 2004a) shows that asymptotically, NM follows a χ^2 distribution with degrees of freedom not dependent on N. This has two immediate consequences. First, $\log(NM)$ and N are asymptotically independent. Second, the log of a χ^2 variable is approximately normal, therefore $\log(NM)$ is approximately normal as well. Due to the independence of $N(t)$ and NM, standard techniques from multivariate analysis are applicable for construction of an elliptical

or other appropriately chosen rejection region for a bivariate normal population at prespecified alpha level that we can use to test for deviations from normalcy.

Another approach we can use is to consider bivariate values in the event of a bioterrorist attack; in this case there is an optimal discriminator (the quadratic classification rule) between two bivariate normal populations (Johnson and Wichern, 2002, Section 11.3) in order to decide if an attack has occurred. This rule defines a classification boundary via a quadratic form (defined by means and covariances of the training set populations) in order to assign new observations to one of the existing populations. The two populations in this case would be the bivariate distribution under the null, and the modeled bivariate distribution under the alternative of a biological attack. The classification rule is a quadratic form that, given $\eta(t)$ and $\log(NM)$ on a particular day, assigns this observation to either the null or alternative population. In Figure 8.5 we illustrate a typical case of the null and alternative populations, together with the boundary of the discriminator. This rule minimizes the expected error of misclassification. The false positive rate can be controlled by shifting the quadratic boundary appropriately, as determined via simulation or resampling of the historical record.

8.4.3 Power Calculations

With each of 1399 days of data, we added a simulated cluster to each day and compared the power of temporal, spatial, and spatio-temporal statistics to

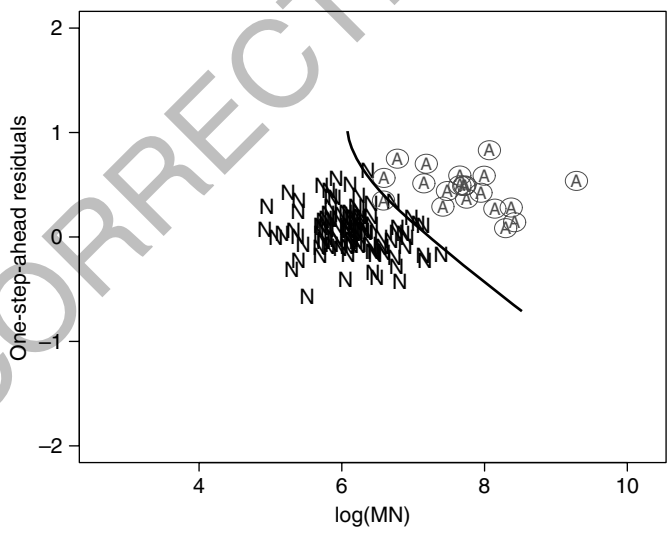

Figure 8.5 Subset of the null (labeled N) and alternative (circled A) populations used to train the quadratic discriminator. A portion of the classification boundary is also shown.

detect such a signal. Power calculations were performed separately for each of the three daily categories, since prediction and behavior differ within each of these categories. We define power as the ratio of daily detections to the total number of days observed. Using the statistics discussed above, we calculated power based on the simulated disease signal in our three constructed data sets.

For the univariate test statistic based on time series modeling, we calculated power to detect a temporal signal in the data. Using the first 1096 days (three full years) to train the model, power was calculated to detect an additional 6, 9, or 12 cases added to the case counts of the final 303 days of data. Results are shown in Table 8.3 (these results are not stratified by location since the statistic depends only on the number of cases and not the spatial locations). We see that the power to detect a disturbance increases as the size of the disturbance increases, as it should.

For the three data sets of clustered data, we calculated values of the two test statistics on each of the 1399 days available, and compared the value of the statistics to their respective distributions under the null hypothesis of no clustering. These null distributions are determined using resampling methods with the unaltered historical data. Results are shown in Table 8.4.

When considering a bivariate statistic we generate a training sample based on a modeled signal consisting of six cases near location 179, superimposed on each of the first 1096 days of data. The temporal test statistic is N, and for the spatial statistic we choose $\log(NM)$. Here, the transformation of the test statistic M to $\log(NM)$ standardizes the distribution for differing numbers of cases, and the logarithm gives a statistic that is roughly normally distributed.

Following this approach, we generate two distinct bivariate normal populations of values, consisting of η residuals together with $\log(NM)$ calculations for 1096 days of null and alternative training data. For the simulated clusters in the final 303 days of data, we calculate the corresponding bivariate test statistic and use the quadratic classification rule to place each day's simulated cluster into the null (no signal) population or the alternative (signal present) population. Power in this case is the number of clusters classified in the alternative over total number of observations. Results are shown in Table 8.4.

The power of the univariate statistic $N(t)$ which detects deviations from the predicted number of cases on a daily basis illustrates the difficulties of time series modeling for public health surveillance. Rather than rely on a simple

Table 8.3 Power to detect temporal clusters.

	Hol./wkends 94 days	Wkdays 165 days	Day after hol. 44 days	Overall weighted average
$N+6$	0.266	0.248	0.250	0.254
$N+9$	0.479	0.315	0.318	0.366
$N+12$	0.755	0.467	0.364	0.541

Table 8.4 Power to detect various cluster models. Group 1 refers to the cluster centered at tract 179, with an additional case added to tracts 179, 182, 183, 184, 191, and 192. Group 2 refers to the cluster centered at tract 477, with an additional case added to tracts 477, 478, 479, 480, 482, and 484.

		Wkends/hol. 438 days	Wkdays 749 days	Day after hol. 212 days	Overall weighted average
Group 1	MEET	0.813	0.194	0.099	0.373
$N+6$	M statistic	0.495	0.362	0.250	0.387
	Bivariate statistic	0.585	0.394	0.227	0.429
Group 2	MEET	0.769	0.085	0.066	0.296
$N+6$	M statistic	0.475	0.295	0.222	0.340
	Bivariate statistic	0.543	0.358	0.250	0.399
Groups 1 & 2	MEET	0.986	0.427	0.226	0.571
$N+12$	M statistic	0.568	0.430	0.325	0.457
	Bivariate statistic	0.904	0.606	0.386	0.667

autoregression, results could be improved by considering a multivariate periodic autoregression (Pagano, 1978). For both the MEET and M statistics, power is consistently higher for weekends and holidays than for other types of days. On weekends and holidays, mean case volume is much lower at the clinics. This leads to a higher signal-to-noise ratio in the simulated data and thus a more detectable spatial aberration when a cluster of fixed magnitude (6 or 12 cases) is added to the data. The MEET is especially sensitive to this type of aberration, as adding one case to a region where the expected number of cases is minuscule greatly inflates the statistic. Although the MEET has especially high power on weekends and holidays, the power of the MEET statistic declines much more rapidly than M as the case volume increases.

Both spatial statistics perform quite well in detecting the simultaneous 179/477 clusters. Superior performance on data sets containing multiple clusters is a characteristic typically shared by distance-based methods of cluster detection as compared to other spatial methods (Kulldorff *et al.*, 2003; Ozonoff *et al.*, 2004).

The bivariate statistic shows promise for an effective use of available data. The power results show that for these simulated clusters, the bivariate approach outperforms the use of purely temporal or purely spatial information.

8.5 LOCATING CLUSTERS

Having decided that the M statistic indicates that there is a deviation from the null distribution, the next step is to determine whether this deviation is caused by an exogenous cluster, or clusters, of individuals, and to locate this cluster or clusters.

There are not too many principled guides in the literature for locating clusters, especially if there is more than one cluster (see Lawson and Denison, 2002, for discussion of small-area data). Fortunately, the M statistic related methods based on distances suggest a natural method for locating clusters.

Here we concentrate on the spatial location problem, leaving the time component to later studies. Let $\{s_i\}_1^n$ be the locations of the individuals, and $\mathbf{D} = (d_{ij})$ be the $n \times n$ distance matrix where d_{ij} is the distance between s_i and s_j. In Figure 8.6 we see a distribution of points in a plane with two clusters of points superimposed. We presume that the null hypothesis about the distribution of $\{s_i\}_1^n$ has been rejected, and we now search to locate the exogenous cluster or clusters that presumably were the reason for rejection.

Consider each row of the matrix of distances, $\mathbf{D}$. Fixing on row i, the $d_{ij}, j = 1, \ldots, n$, are a sample of independent distances from the point s_i. From the null distribution, either theoretically or via Monte Carlo, we can determine what the null distribution of points from s_i should be. Then we can compare this distribution with the observed distribution of distances from s_i, and presumably will be able to discern the points s_j that belong to an exogenous cluster. Of course, it is too much to hope that for a single i we will pick these s_j with any confidence, but if we gather information from all the s_i, one at a time, we can use the aggregate information to identify the clusters. This is the intuitive description of the method we use.

Choose a row $i, i = 1, \ldots, n$, and determine the null distribution of the distances from s_i. This may have to be achieved by resampling points from the null distribution of points. Having determined this distribution, then, for a fixed

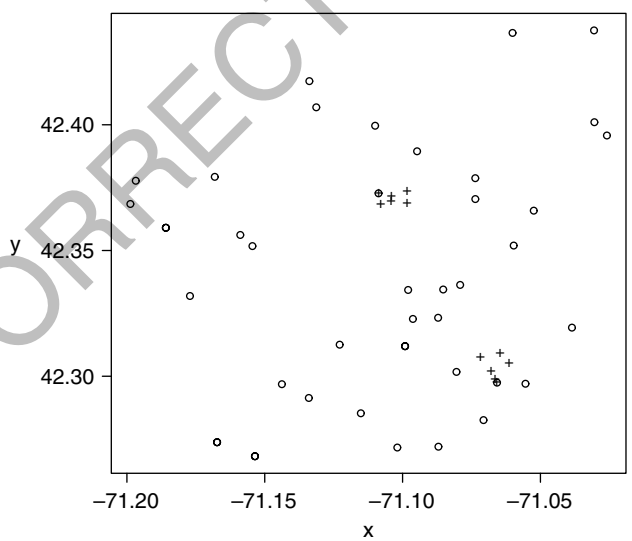

Figure 8.6 Typical distribution of cases for the Massachusetts data set. Superimposed clusters are denoted by a '+'.

integer $k > 1$, determine the k equispaced quantiles for the distribution, and hence create k equiprobable bins to receive the d_{ij}. The d_{ij} associated with the s_j in exogenous clusters will presumably give rise to bins with excessive counts. These s_j from exogenous clusters will presumably have a similar impact for other i, and so a record can be kept of the s_j which appear in bins that are oversubscribed, as we consider each row i.

To aggregate over the rows, consider a scoring system. For each row i, let $\text{score}(i,\ j) = 1$ if s_j belongs to an oversubscribed bin. Then with each point s_j associate the

$$\text{score}(j) = \sum_{i=1}^{n} \text{score}(i, j).$$

Subsequently look at these scores to determine which ones are too large. These are the ones that can be tagged as belonging to the exogenous clusters.

The binning process described above is a traditional way of determining goodness of fit. One of its disadvantages is that the underlying distribution of points is continuous, whereas the binning is inherently discrete. This may manifest itself in points which are in an exogenous cluster but, because of the discrete character of the bins, fall just next door to an oversubscribed bin. To overcome this effect of discretization, we compromise by defining a score function which takes the value one for an oversubscribed bin, and the value 0.50 for the bins on either side of the oversubscribed bin. If the oversubscribed bin is on a boundary (either it is the first or last bin) then it will only have one neighbor.

This scoring system is, of course, one of an infinite number of scoring systems one can devise.

The only remaining unknown is the definition of what we mean by an oversubscribed bin. For a fixed i we can consider the n distances d_{ij} as a sample of independent and identically distributed variables. Thus the counts of the numbers falling into the k bins can be considered as the realization of a multinomial distribution of size n with equiprobable cells, each with probability $1/k$. We can determine a bin to be oversubscribed if the number of distances in the bin exceeds n/k by two standard deviations. Other cutoffs can be entertained.

The last step is then to determine how large $\text{score}(j)$ must be before we declare s_j to be a location within a cluster. A cutoff can be determined via Monte Carlo methods.

This method is exemplified below.

We now return to the example taken from data from a large health care provider in Massachusetts. We wish to show the efficacy of the above method in locating clusters in the data. Again, we subset the data by the day of the week (weekends/holidays, day after weekend/holiday, and weekdays).

As a measure of the adequacy of this method, we borrow from methods used in diagnostic testing and report estimates of sensitivity and specificity. Suppose our method tags b regions as a comprising cluster or clusters. In our setting,

we define sensitivity as the probability of detecting the regions that actually constitute the cluster(s). Specificity is defined as the probability that the regions that are not tagged are not in the cluster(s).

Table 8.5 provides a summary of the results of this method applied to the Massachusetts data. Here we give results for detecting the three cluster models (region 179, 477, and a cluster in 179 and 477 simultaneously) for the three different types of days (weekends and holidays, weekdays, and days after holidays). We consider three different significance levels for determining the cutoff for the scores: 0.05, 0.10 and 0.15. Increasing the significance has the effect of increasing sensitivity and decreasing specificity. However, specificity remains high in all scenarios.

In the surveillance setting, we would often be satisfied with detecting at least part of the cluster. Therefore, we can imagine a much more forgiving definition of sensitivity as the probability of detecting at least one of the cluster regions. In other words, we are not concerned that we detect all of the regions in the cluster, as health professionals alerted by the alarm would likely fan out from investigating that region to surrounding areas that would likely comprise the cluster. Were we to use this as a measure of efficacy, the method would undoubtedly appear even better. The last column probably best approximates the ubiquitous 95 % specificity.

On the other hand, a high specificity is also desirable. A typical system may require a decision each day and missing an outbreak might lead to disastrous consequences; but, by the same token, too many false positive alarms might lull the analyst into treating the system with skepticism and subsequently missing a valid alarm. It is thus comforting to see the high specificities in Table 8.5.

Table 8.5 Sensitivities and specificities for locating the clusters with the Massachusetts data set. Results are given with the cutoff for the score being determined as the 95th, 90th or 85th percentiles of the empirical distribution of the scores.

	95th		90th		85th	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
Group 1, $N+6$						
Hol./wkends	0.76	0.99	0.84	0.99	0.89	0.99
Wkdays	0.61	0.98	0.73	0.96	0.89	0.95
Day after hol.	0.59	0.97	0.68	0.95	0.79	0.94
Group 2, $N+6$						
Hol./wkends	0.77	0.99	0.84	0.99	0.88	0.98
Wkdays	0.60	0.98	0.75	0.96	0.82	0.98
Day after hol.	0.63	0.97	0.72	0.95	0.81	0.94
Group 1 and 2, $N+12$						
Hol./wkends	0.45	0.99	0.61	0.99	0.76	0.98
Wkdays	0.45	0.98	0.62	0.96	0.72	0.95
Day after hol.	0.48	0.97	0.63	0.95	0.73	0.93

8.6 CONCLUSION

Spatial surveillance has as it goal the recognition of deviations from the 'normal' distribution of events in a region. We have shown the utility of distance methods in achieving the stated objectives of spatial surveillance. Distance methods are characterized as statistical methods that utilize the distances between events in detecting aberrations in spatial behavior.

Two statistics are immediately applicable for use in spatial surveillance. The MEET statistic is widely recognized in contemporary literature and practice. It is only applicable to aggregated data, such that the data consists of counts of events within each spatial region. The statistic has been shown to have good power, especially when case volume is low relative to the increase in the number of cases attributed to a cluster.

Much of the focus of this chapter has been on the M statistic. This statistic seeks to detect changes in the distribution of the interpoint distances by considering a statistic that is similar to the Mahalanobis distance. This can be easily applied to data streams with either aggregated or exact location information.

The M statistic can further be extended to incorporate temporal trends in the data stream, as exemplified by the bivariate statistic illustrated above. Not surprisingly, this leads to an increase in power for the detection of anomalies in the data, as one would expect such a disturbance to represent an increase in case volume, as well as a disturbance in the spatial distribution.

Spatial surveillance requires not only an alarm to be sounded when a disturbance has occurred, but also some indication of the location and shape of the disturbance to facilitate further investigation and efficient methods to control and diffuse the source of the disturbance, inhibiting further spread to the population. We have shown an effective method for locating the source of the signal causing an alarm with the M statistic. Such methods are crucial to the success and efficacy of a surveillance system.

Distance methods are a natural tool for spatial surveillance. The issues presented by this problem require methods that incorporate information extending beyond a simple mean or other traditional statistics that are often employed when confronted with data on the real line. The increased dimensionality and correlation of the data call for methods that can distinguish between normal and abnormal behavior for an infinite number of scenarios that are not easily characterized or classified. Distance methods appear to have the potential to capture the complexity of a spatial distribution. Further, statistics such as M allow for incorporation of additional information into the data stream, such as temporal trends.

It would be unfair to fail to recognize the many difficulties that arise when working with distance data. As has been illustrated above, these methods still require much refinement and further research. Working with the dependencies intrinsic in interpoint distances is complicated and requires further rigorous investigation. Much of these complexities can be circumvented in practical

implementation via resampling methods. However, to better understand, generalize, and optimize these statistics, greater theoretical understanding is needed. It has also been shown that theoretical developments can lead to an increase in efficiency and decreases in computation time. This is exemplified by the estimation of the variance–covariance matrix for the *M* statistic.

We advocate the use and further development of distance methods in spatial surveillance. These methods have been shown to be effective and complementary when compared to quadrat methods, such as the spatial scan statistic. Further, the *M* statistic has great promise in detecting spatial aberrations that extend beyond simple circular clusters. Ideally, a surveillance system would make use of multiple statistical methods, coupled with vigilant and timely epidemiological investigation of alarms raised by these automated methods.

ACKNOWLEDGMENTS

This research was partially funded by grants from the National Institutes of Health, RO1AI28076, and the National Library of Medicine, RO1LM007677.

REFERENCES

Alexander, F.E. and Boyle, P. (eds) (1996) *Methods for Investigating Localized Clustering of Disease*. Lyon: International Agency for Research on Cancer.

Bartlett, M.S. (1964) The spectral analysis of two-dimensional point processes. *Biometrika*, **51**, 299–311.

M. Bonetti and M. Pagano. (2004a) The interpoint distance distribution as a descriptor of point patterns, with an application to cluster detection. *Statistics in Medicine*. In press.

M. Bonetti and M. Pagano. (2004b) Parametric estimation of interpoint distance distributions, with an application to biosurveillance data. *Biometrika*. Submitted.

Brookmeier, R. and Stroup, D. (eds) (2004) *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health*. Oxford: Oxford University Press.

Couteron, P. and Kokou, K. (1997) Woody vegetation spatial patterns in a semi-arid savanna of Burkina Faso, West Africa. *Plant Ecology*, **132**, 211–227.

Diggle, P.J. and Chetwynd, A.G. (1991) Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, **47**, 1155–1163.

Dobbertin, M., Baltensweiler, A. and Rigling, D. (2001) Tree mortality in an unmanaged mountain pine (*Pinus mugo* var. *uncinata*) stand in the Swiss National Park impacted by root rot fungi. *Forest Ecology and Management*, **145**, 79–89.

Elliott, P., Wakefield, J., Best, N. and Briggs, D. (2000) *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford University Press.

Johnson, R.A. and Wichern, D.W. (2002) *Applied Multivariate Statistical Analysis* (5th edn). Upper Saddle River, NJ: Prentice Hall.

Kulldorff, M. (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

Kulldorff, M. (1998) Statistical methods for spatial epidemiology: tests for randomness. In M. Löytönen and A. Gatrell (eds), *GIS and Health*, pp. 49–62. London: Taylor & Francis.

Kulldorff, M., Tango, T. and Park, P. (2003) Power comparisons for disease clustering tests. *Computational Statistics and Data Analyis*, **42**, 665–684.

Lawson, A.B. (2001) *Statistical Methods in Spatial Epidemiology*. Chichester: John Wiley & Sons, Ltd.

Lawson, A.B. and Denison, D. (2002) *Spatial Cluster Modelling*. Boca Raton, FL: Chapman & Hall/CRC.

Ozonoff, A., Bonetti, M., Forsberg, L. and Pagano, M. (2004) Power comparisons for an improved disease clustering test. *Computational Statistics and Data Analyis*. To appear.

Pagano, M. (1978) On periodic and multiple autoregressions. *Annals of Statistics*, **6**, 1310–1317.

Rao, C.R. and Mitra, S.K. (1971) *Generalized Inverse of Matrices and its Applications*. Wiley.

Reis, B.Y., Pagano, M. and Mandl, K.D. (2003) Using temporal context to improve biosurveillance. *Proceedings of the National Academy of Sciences of the USA*, **100**, 1961–1965.

Ripley, B.D. (1976) The second-order analysis of a stationary point process. *Journal of Applied Probability*, **13**, 255–266.

Silverman, B.W. (1976) Limit theorems for dissociated random variables. *Advances in Applied Probability*, **8**, 806–819.

Tango, T. (1995) A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine*, **14**, 2323–2334.

Tango, T. (2000) A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine*, **19**, 191–204.

Whittemore, A.S., Friend, N., Brown, B.W. and Holly, E.A. (1987) A test to detect clusters of disease. *Biometrika*, **74**, 631–635.