Change-point Problems
IMS Lecture Notes – Monograph Series (Volume 23, 1994)

DETECTING CLUSTERS IN DISEASE INCIDENCE

BY DANIEL RABINOWITZ
Harvard School of Public Health

This paper is concerned with searching for localized environmental risk factors. The approach taken here uses case-control data to search for clusters of disease cases. In this context, case-control data means a sample of locations associated with diseased subjects (cases) and healthy subjects (controls). A cluster of cases is a region where the number of cases appears to be larger than what would have been expected had the cases occurred randomly in the underlying population. Clusters indicate areas where localized risk factors are likely. The methodology developed here produces a random field over the region where the cases and controls are located. The field is large where there are clusters of cases. Asymptotically, as the number of cases and controls becomes large, the field tends in distribution to a smooth Gaussian field. The operating characteristics of inferential procedures based on the random field may be approximated by considering the random field's limiting distribution.

1. Introduction. Environmental risk factors such as toxic spills, contaminated drinking water and radiation may increase the incidence of cases of birth defects, cancer or disease. When exposure to a risk factor occurs in small areas or during small periods of time, the increased incidence of cases may take the form of a spatial or temporal cluster. If a cluster of cases is detected, health workers can scrutinize the location of the cluster and, one hopes, discover and eliminate localized risk factors that may have caused the increased incidence. Decisions to further investigate locations identified by a cluster detection methodology must be made with reference to the probability of incorrectly detecting a cluster where there is no increased risk.

Case-control data is a sample of locations associated with diseased subjects (cases) and a sample of locations associated with healthy subjects (controls). This paper presents a method for using case-control data to detect clusters of cases. The control data is used to account for non-homogeneities in the density of the population from which the cases arise as suggested in

AMS 1991 Subject Classification: Primary 60G70; Secondary 62P99

Key words and phrases: Cluster, point process, screening, scan statistic, tail probability.

Cuzick and Edwards (1990). An approach to approximating the probability of detecting clusters when there are no areas of increased risk is also presented.

Cluster analyses are undertaken with various purposes. In some situations, it is of interest to test whether there is a cluster near some posited source of risk. An example is Diggle (1990). In some situations it is of interest to determine whether cases tend to arise in clusters as opposed to being spread out uniformly. Examples are Tango (1984), Nagelerke et al. (1988) and Fraser (1983). In other situations, the purpose is to find clusters of cases that may indicate exposure to localized risk factors. Examples are Besag and Newell (1991) and Oppenshaw et al. (1987). It is the third purpose, finding clusters, that motivates the methods explored here.

In searching for clusters of cases for the purpose of discovering localized environmental risk factors, a main concern is to insure that apparent clusters are not simply artifacts of variations in the spatial or temporal density of the underlying population from which the cases arise. Areas where the underlying population is dense will contain more cases than will areas where the underlying population is sparse. If non-homogeneities in the density of the population are not accounted for, then these areas where the incidence of cases is high may be mistaken for areas of increased risk.

Two approaches are generally taken in response to the concern. One is to use controls or historical data to form an estimate of the the density of the underlying population, and then to compare the locations of the cases to the estimated density. Examples are Diggle (1990) and Ohno et al. (1979). Another is to model the density of the population at risk as a nuisance parameter, and to take as the reference distributions to which test statistics are compared, the test statistics' conditional distributions given sufficient statistics for the nuisance. Examples are Whittemore (1987) and Weinstock (1981). The conditioning approach, which is taken here, sidesteps the difficulties inherent in estimating the density of the underlying population.

Often, cluster analyses are based on a discretization. The spatial or temporal region where the cases and controls are located is partitioned, more or less arbitrarily, into disjoint units. An unusually large number of cases in a unit or in adjacent units is considered evidence of localized increased risk. Examples are Raubertas (1988) and Ederer et al. (1964). Ederer points out the importance of choosing the partition without reference to the data.

An alternative approach, based on scan statistics, takes into account the fact that clusters are not likely to occur in arbitrarily delimited units. In a scan statistic approach, a statistic chosen to be sensitive to a local increase in the incidence of cases is evaluated at each point of the spatial or temporal region of interest. The function that maps each point of the region to its associated value of the test statistic forms a random field over the region. Points where the field takes on extreme values are considered locations where there may be localized risk factors. An example is Wallenstein (1980). The methodology developed here is a scan statistic approach.

All cluster analyses must grapple with the question of what sort of clusters are envisioned as likely. In scan statistic approaches, the question manifests itself as the choice of the scan statistic that is to be computed at each point of the spatial or temporal region. In approaches based on discretization, the question manifests itself as the problem of choosing the unit size and of how the information in nearby units is to be combined. The methodology explored here is flexible enough to accommodate various degrees of knowledge about the sorts of clusters likely to be induced by risk factors of interest.

A scan statistic approach is an example of multiple testing; the hypothesis that there is a cluster located at any given point is tested at the infinitude of points of the spatial or temporal region of interest. Deciding what sort of p-value should be associated with the tests or how to compute a p-value can be a formidable problem. To the policy maker faced with the decision of whether or not to expend resources examining area pointed out by an extreme value of the random field, the probability that a cluster will be detected where there are no localized risk factors is an important quantity. On the other hand, to someone at a location where excess risk is indicated by an extreme value of the random field, it is irrelevant that tests were performed at other locations. In any case, it is useful to understand the probabilistic behavior of the random field that results from a scan statistic approach.

In order to evaluate the statistical significance of an extreme value of the scan statistic achieved at some point in the region, it is often suggested that the the extreme value be compared to the distribution (under the hypothesis of no localized risk factors) of the maximum of the random field. In a few cases, especially ones in which the scan statistic corresponds to the number of cases in a fixed window of time, exact or asymptotic formulae for the distribution of the maximum have been derived. Examples are Huntington and Naus (1975), Gates and Westcott (1985), Glaz (1989) and Berman and Eagleson (1985). Loader (1991) treats some more complicated situations. Sometimes it is suggested that distribution theory be determined via Monte Carlo simulation. An example is Day et al. (1988). Here, an approach based on approximating the distribution of the random field by the distribution of a Gaussian field with the same covariance structure is used to derive a formula for tail probabilities of the maximum.

In the next section, the kind of case-control data to which the methodology is applicable is described and a model for the data, under the hypothesis of no localized risk factors, is proposed. The model is then extended to a family of alternatives which are used to suggest a scan statistic. The approximation to the tail probabilities for the maximum of the random function that results when the scan statistic is computed at each point of the region of interest is then presented. The third section contains a heuristic derivation of the approximation to the distribution of the maximum. In the fourth section, the methodology is applied to some real data and the results of some simulation experiments are presented.

2. Detecting Clusters. This section begins with a description of the sampling scheme assumed for the case-control data. Then, a non-parametric model for the locations of the cases and controls under the null hypothesis of no localized risk factors is posited. In order to develop a statistic sensitive to a localized increased in the incidence of of cases, the model for the null hypothesis is embedded in a semi-parametric family of alternatives reflecting localized increased risk, and the model is used to derive a likelihood-based score statistic. A normalized version of the score is suggested for use as a scan statistic. When the scan statistic is evaluated at each point of the region of interest, a random field results. Large values of the field indicate points where there is an increased incidence of cases and where increased risk is likely. This section concludes with a description of an approximation to tail probabilities for the maximum of the field. The approximation may be used to assess the statistical significance of large values of the field.

The methodology developed here requires a random sample of the locations of cases and controls. Cases are subjects with a disease that may be associated with localized environmental risk factors. Controls are healthy subjects drawn from the same underlying population as the cases. Because the controls are used to adjust for non-homogeneities in the density of the underlying population, if there are no localized risk factors, the density of the locations of the controls must be the same as the density of the locations of the cases. If the cases and controls are located in a spatial domain, the locations take the form of two dimensional vectors. If the cases and controls occur in space and time, then the locations take the form of three dimensional vectors.

Let m denote the number of cases and let n denote the number of controls. Denote the locations in the combined sample of cases and controls by $x_1, x_2, \ldots, x_{n+m}$. A probabilistic model for the sampling scheme, under the hypothesis of no localized risk factors, is that each subset of the combined sample that contains m elements is equally likely to be the subset associated with the cases. That is, given the number of cases and controls, m and n, and given the locations of the combined sample of cases and controls, $x_1, x_2, \ldots, x_{n+m}$, the subset of locations associated with cases behaves like a random subset of $\{x_1, x_2, \ldots, x_{n+m}\}$ drawn uniformly from all subsets of size m. In what follows, this model will be referred to as H_0.

In order to develop test statistics, it is convenient to embed the null hypothesis, H_0, in a family of models similar to the models considered in Diggle (1990). Suppose that, in the absence of localized risk, the locations of the cases and controls are generated by two Poisson processes. Suppose also that the rate for the process generating the locations of the cases is $\rho\lambda(x)$ and that the rate for controls is $\lambda(x)$. Here, the argument x takes values in the whole spatial or temporal region of interest. The unknown function λ reflects non-homogeneity in the spatial or temporal distribution of the population at risk. The unknown constant ρ reflects that, although the controls arise from the same population at risk as the cases, they may not arise with the same frequency.

In order to use this Poisson process model to develop a test statistic sensitive to localized increased incidence of cases, the model is extended to a family of alternatives that correspond to increased risk around a given location, say t. Consider alternatives, indexed by β, where the rate for the process generating cases is replaced by a rate of the form $\rho\lambda(x)e^{\beta g(x,t)}$. Here $g(x,t)$ is a known function that reflects the distance between locations x and t. If $g(x,t)$ were large when x is near t, and small for x away from t, then for positive β, the ratio of the rate for cases to the rate for controls would be larger for x near t than for x away from t. The larger that β is, the more extreme would be the differences in the ratio. In this way, β would parameterize a family of alternatives that reflect increased risk near t. The model is semi-parametric in the sense that λ is an infinite dimensional nuisance parameter while β is a finite dimensional parameter of interest.

Let z_i, $i = 1, 2, \ldots, z_{n+m}$, denote the indicator that the i^{th} location is associated with a case, so that z_i is equal to 1 if the i^{th} location is associated with a case and z_i is equal to 0 otherwise. Under the hypothesis, $\beta = 0$, the conditional distribution of the z_i, given the x's, n, and m is the same as the the distribution associated with the null hypothesis, H_0. That is, all subsets of cases are equally likely. In this sense, H_0 is embedded in the family of alternatives.

The scan statistic advocated here may be derived by considering the problem of testing whether $\beta = 0$ in the model for increased risk. The likelihood for the semi-parametric model is given by

$$\exp\left\{-\int \lambda(x)\left(1+\rho e^{\beta g(x,t)}\right)dx\right\}\prod_{i=1}^{n+m}\left(\lambda(x)\rho e^{\beta g(x,t)}\right)^{z_i}\lambda(x)^{(1-z_i)}.$$

The conditional likelihood of the z's, given the x's, n and m is

$$\frac{\prod_{i=1}^{n+m} e^{z_i \beta g(x_i,t)}}{\sum_{w \,\epsilon\, S} \prod_{i=1}^{n+m} e^{1_{\{i \,\epsilon\, w\}} \beta g(x_i,t)}},$$

where S is the class of subsets of $\{1, 2, \ldots, m+n\}$ of size m. Inference for β may be based on the score from the conditional likelihood:

$$T_t \;=\; \frac{d}{dt}\ell_t(\beta)\bigg|_{\beta=0} \;=\; \sum_{i=1}^{n+m} (z_i - p)(g(x_i,t) - \bar{g}(t))$$

where p is the proportion of cases in the combined sample of cases and controls, $m/m+n$, and $\bar{g}(t)$ is the sample average of the g's,

$$\frac{1}{m+n} \sum_{i=1}^{n+m} g(x_i, t).$$

The following argument suggests that when g is chosen so that the family of alternatives reflects increased risk around a point t, the score from the conditional likelihood is sensitive to an increased incidence of cases near to t. Consider again the situation in which $g(x,t)$ is large when x is close to t, and small when x is far from t. Suppose that the proportion of cases at the locations near t is larger than the proportion of cases at the locations away from t. Then, $z_i - p$ would be, on average, greater than zero among those i for which $g(x_i,t) - \bar{g}(t)$ is positive, and, on average, less than zero among those i for which $g(x_i,t) - \bar{g}(t)$ is negative. It follows that in such situations, T_t would be large and positive. Similarly, if the proportion of cases at the locations near to t is close to the proportion of cases at the locations away from t, then T_t is likely to be around 0. In this way, the value of T_t indicates whether subjects near to t appear to be at excess risk for becoming a case. As will be explained, the scan statistic advocated here is a normalized version of T_t.

If attention is restricted to the conditional likelihood, then there is no need to estimate λ, the density of the population from which the cases and controls arise. Furthermore, the score from the conditional likelihood has an optimality property in the context of the semi-parametric model. In the asymptotic scenarios in which, as N tends to infinity, λ takes the form $N\lambda_0$ while ρ remains fixed, the conditional maximum likelihood estimator, given by solving the score equation

$$\frac{d}{dt}\ell_t(\beta) = 0$$

for β, has a variance equal to the asymptotic variance of the maximum likelihood estimate of β in a particular parametric sub-model. The import of this result is that, in scenarios in which the numbers of cases and controls are both large, and in which there is no auxiliary information about the density of the underlying population from which cases and controls arise, restricting attention to the conditional likelihood does not lead to any loss of asymptotic efficiency.

The parametric sub-model alluded to above may be associated with the likelihood, parameterized by α, β and ρ,

$$\exp\left\{-\int N\lambda_0(x)e^{\alpha h(x,t)}\left(1+\rho e^{\beta g(x,t)}\right)dx\right\}\prod_{i=1}^{n+m} N\lambda_0(x_i)e^{\alpha h(x_i,t)}\left(\rho e^{\beta g(x_i,t)}\right)^{z_i},$$

where

$$h(x,t) = g(x,t) - \mu(t)$$

and where

$$\mu(t) = \frac{\int \lambda_0(x)(1+\rho)g(x,t)dx}{\int \lambda_0(x)(1+\rho)dx}.$$

Let ν equal

$$\frac{\rho}{(1+\rho)^2}\int \lambda(x)(1+\rho)(g(x,t)-\mu(t))^2dx.$$

Then the information in the parametric sub-model evaluated at α and β both equal to 0 is $n\nu$. Observe that at $\beta = 0$, the information in the conditional likelihood is

$$\sigma(t) = p(1-p)(1-\delta)\sum_{i=1}^{n+m}(g(x_i,t)-\bar{g}(t))^2,$$

where $\delta = 1/(n+m-1)$. That the conditional likelihood estimator is asymptotically as efficient as the maximum likelihood estimator in the parametric sub-model follows therefore from a law of large numbers result: as N tends to infinity, $\sigma(t)/n\nu$ tends to 1 in probability.

Now, consider the function g. It should be chosen to be sensitive to the kinds of clusters that are envisioned as probable. If it is thought that likely environmental risk factors produce the disease in question only in nearby subjects, then $g(x,t)$ should be large for x close to t and should decrease quickly as x moves away from t. If it is thought that the probability of the disease is raised for subjects only moderately close to a risk factor, then $g(x,t)$ should decrease slowly as x moves away from t.

When there is no clear notion of what kinds of clusters are probable, then T_t might be evaluated for several choices of g. In fact, the test statistic could be evaluated as as the g function ranges over a continuum of possibilities, and the supremum of normalized versions of the resulting values might be used as a test statistic.

The g function or functions might also be chosen to reflect characteristics of the region known to influence how exposure to a risk factor at one location, say t, might be experienced by a subject at another location, say x. Geographic distance may not be adequate for expressing the influence that a localized risk factor has on the likelihood of disease at nearby locations. For example, even though x and t might be close, if they were separated by a a mountain range that blocks the spread of airborne toxic materials, then they might be appropriately treated as far apart. On a smaller scale, if the location x were contained in some neighborhood near to the location t, but if the inhabitants of the neighborhood were thought not to spend time in the area around t, then it could be appropriate to treat x and t as far apart.

The derivation in the next section will rely on the assumption that g is smooth. In the examples of the fourth section, $g(x,t)$ is of the form $e^{-\eta\|x-t\|^4}/(1+e^{-\eta\|x-t\|^4})$.

Now, consider how T_t may be used to define a score statistic that may be evaluated at every point of the spatial or temporal region of interest. (Strictly, it is not possible to evaluate a statistic everywhere in the region, but the test statistic may be evaluated over a fine grid of locations.) It is desirable that values of the test statistics computed at different locations be roughly comparable. If they were, then points where localized risk factors are most likely to be found could be determined by finding the largest values of the random field. The statistic T_t has the disadvantage that at different values of t, the conditional variance, $\sigma(t)$, is not the same. This suggests that it would be easier to compare evidence of excess risk at different locations using the normalized version of T_t, $W_t = \sigma^{-\frac{1}{2}}(t)T_t$. The variance of W_t is equal to 1 at every t. This makes values of the test statistic evaluated at distinct locations at least roughly comparable.

When T_t is evaluated at only one g function, points where W_t is large are to be considered locations where localized risk factors are likely, and worthy of further investigation. When T_t is evaluated over a continuum of g functions, indexed, say by η, points where the maximum (as η varies) of W_t are to be considered locations where localized risk factors are likely. Alternatively, since the form of T_t is reminiscent of a kernel density estimate, the parameter η might play a role similar to that of a band width and η might be chosen subjectively after an inspection of the field.

For large n and m, under broad regularity conditions on g and the distribution of the x_i, under H_0, the conditional distribution, given the x_i, n and m, of W_t evaluated at a particular point t, is approximately that of a standard normal. The conditional covariance of W_t evaluated at two points, t_1 and t_2, is

$$\sigma_{t_1,t_2} = \sigma^{-\frac{1}{2}}(t_1)\,\sigma^{-\frac{1}{2}}(t_2)\sum_{i=1}^{n+m}(g(x_i,t_1)-\bar{g}(t_1))(g(x_i,t_2)-\bar{g}(t_2))p(1-p)(1-\delta)$$

where $\delta = 1 - 1/(n+m-1)$. As t varies, W_t forms a random field. Under broad regularity conditions, the conditional distribution of W_t, as a field, may be approximated by the distribution of a smooth Gaussian field with covariance σ_{t_1,t_2}.

The decision to investigate further a cluster indicated by a large value of W_t must depend in part on the statistical significance of the detected cluster. In order to evaluate how unusual extreme values of W_t are, it is useful to have an approximation to the conditional distribution of the maximum of the field.

Denote by M, $\sup_t W_t$, the maximum of the field. An approximation to the tails of the distribution of M based on the Gaussian field approximation is

$$P\{M > b\} \sim \gamma_b = \int_A (2\pi)^{-\frac{d}{2}} b^{d-1}\varphi(b)\,|\Lambda_t|^{\frac{1}{2}}\,dt + \frac{1}{2}\int_{\partial A} b^{d-2}(2\pi)^{-\frac{d-1}{2}}\varphi(b)\,|\Lambda_t'|^{\frac{1}{2}}\,dt$$

where $\varphi(b)$ is the density of a standard normal; d is the dimension of the spatial or temporal region of interest; A is the region and ∂A is its (smooth) boundary; $-\Lambda_t$ is the matrix of mixed partial derivatives with respect to s of $\sigma_{s,t}$, evaluated at $s = t$; $|\Lambda_t|$ is its determinant; and $|\Lambda_t'|$ is the determinant of $P_t^T \Lambda_t P_t$ where P_t is a d by $d-1$ matrix comprised of orthonormal vectors orthogonal to n_t, a vector normal to the tangent to ∂A at t.

The accuracy of the approximation is sensitive to the extent to which the marginal distribution of W_t is close to Gaussian. When g or the configuration of the x_i is such that for some values of t only a few locations contribute substantially to W_t, or when n and m are small, the Gaussian approximation may be poor. In such situations, simulations of the conditional distribution of the field, although time consuming, might be more appropriate than relying on the approximation. The integrals in the formula for γ_b may be computed approximately by sums.

When the scan statistic is to be evaluated over a continuum of g functions, say indexed by η, as well as over the spatial or temporal region of interest, some modifications must be made to the approximation. The region, A, should be replaced by the set of (t,η) pairs for which the statistic is evaluated and the boundary of A should be replaced by the boundary of the (t,η) pairs; d

should be replaced by $d + d'$, where d' is the dimension of η; and $-\Lambda_t$ should be replaced by the matrix of mixed partials with respect to t and η.

The approximation to tail probabilities for M is appropriate as the distribution of W_t approaches the distribution of a smooth Gaussian field with a covariance which is of the form, for h small,

$$\sigma_{t,t+h} = 1 - h^T \Lambda_t h + O(h^3).$$

In the third section, the approximation is developed heuristically. At the end of the third section, a modification to the formula, applicable when the supremum is taken over only a finite grid of values of the field rather than over the whole region, is described.

Finally, for computational purposes, it is useful to note that the (i_1, i_2)th entry of Λ_t may be written as

$$\frac{B_{i_1}(t)B_{i_2}(t)}{\sigma^2(t)} - \frac{C_{i_1,i_2}(t)}{\sigma(t)}$$

where

$$B_{i_1}(t) = \sum_{i=1}^{n+m} \left(\frac{\partial}{\partial t_{i_1}} g(x_i, t) - \frac{\partial}{\partial t_{i_1}} \bar{g}(t) \right) (g(x_i, t) - \bar{g}(t)),$$

and where

$$C_{i_1,i_2}(t) = \sum_{i=1}^{n+m} \left(\frac{\partial}{\partial t_{i_1}} g(x_i, t) - \frac{\partial}{\partial t_{i_1}} \bar{g}(t) \right) \left(\frac{\partial}{\partial t_{i_2}} g(x_i, t) - \frac{\partial}{\partial t_{i_2}} \bar{g}(t) \right).$$

Here, $\frac{\partial}{\partial t_i}$ means differentiation with respect to the i^{th} component of t.

3.The Approximation. In this section an approach to deriving the approximation of the tail probabilities of the supremum over t of W_t, $P\{M > b\} \sim \gamma_b$, is discussed heuristically. There are three steps to the derivation. The first step is the approximation of the distribution of the supremum over t of W_t by the distribution of the supremum over t of $\widetilde{W}_t$, where $\widetilde{W}$ is a smooth Gaussian field with the same covariance structure as W. The second step is the approximation of the probability that there exists a local maximum of $\widetilde{W}$ exceeding b, for large b, by the expected number of such maxima. The third step is the approximation of the expected number of local maxima exceeding b, for large b, by γ_b. The validity of this approach follows from the fact that $\widetilde{W}$ exceeds b if and only if there exists a local maximum of $\widetilde{W}$ exceeding b.

An approach based on differential geometric arguments to approximating the distribution of the maximum of smooth Gaussian fields is explored in Knowles and Siegmund (1989), Sun (1990), Johnstone and Siegmund (1989) and the references therein. The approach taken here rests on the notion, suggested in Aldous (1989), that for large b the point process of local maxima exceeding b behaves approximately like a Poisson process with a low intensity. The goal of this section is a presentation of the proof idea rather than a rigorous treatment.

For the first step in the derivation, note that the validity of the approximation of the distribution of the supremum of W_t by the distribution of the supremum of a Gaussian field with the same covariance structure would follow from the continuous mapping theorem if weak convergence of W to $\widetilde{W}$ (with respect to the uniform metric) could be established. In suitably regular asymptotic scenarios in which the number of cases and controls tend to infinity on the same order, convergence of the finite dimensional marginals might be demonstrated by an adaptation of the methods of Holst (1979) of Rosén (1972). Tightness might be approached with the methods of Bickel and Wichura (1971). In what follows, it is taken for granted that the distribution of the supremum of W is well approximated by the distribution of the supremum of $\widetilde{W}$.

Consider next the third step of the derivation, the approximation of the expected number of local maxima of $\widetilde{W}$ exceeding b by γ_b. The local maxima may be divided into two subsets, local maxima in the interior of the region A and local maxima on the regions boundary ∂A. The two terms in the definition of γ_b correspond to approximations to the expected number of maxima in each subset.

A local maximum of $\widetilde{W}$ occurs at a point t on the boundary if and only if the following two conditions hold: as a function whose domain is restricted to the boundary of A, $\widetilde{W}$ achieves a local maximum at t; and along paths that pass from the interior of A, through t, to the exterior of A, $\widetilde{W}$ is increasing at t. Given that $\widetilde{W}$, as a function whose domain is restricted to the boundary of A, achieves a local maximum greater than b at a point t, the conditional probability that $\widetilde{W}$ is increasing on paths that pass from the interior of A, through t, to the exterior of A is $1/2$. This suggests that the expected number of local maxima of $\widetilde{W}$ exceeding b, for $\widetilde{W}$ as a function whose domain is restricted to the boundary of A, is twice the expected number of local maxima of $\widetilde{W}$ exceeding b occurring on the boundary.

The first term of γ_b is the integral over the interior of A of an approximation to to the intensity of the point process of local maxima of $\widetilde{W}$ exceeding b. The second term is $1/2$ times the integral over the boundary of A of an approximation to the intensity of the point process of local maxima of $\widetilde{W}$ exceeding b for $\widetilde{W}$ as a function whose domain is restricted to the boundary of A. The derivations of the approximations to the intensities of the two point processes of local maxima are similar, so only the approximation of the process on the interior is considered here.

Every local maximum of $\widetilde{W}_t$ exceeding b in the interior of A is a point where $\widetilde{W}$ is greater than b and where the gradient of $\widetilde{W}$, $\nabla\widetilde{W}$, is equal to 0. Not every such point is a local maximum because of the possibility of saddle-points and local minima. But, as will become apparent, for large b, points where $\widetilde{W}$ is greater than b and where $\nabla\widetilde{W}_t$ is equal to 0 are points where the the matrix of mixed partial derivatives of $\widetilde{W}_t$ with respect to t, $D^2\widetilde{W}_t$, is with high probability negative definite. This suggests that in order to approximate the intensity of the point process of local maxima of $\widetilde{W}$ exceeding b, it suffices to approximate the intensity of the point process of locations where $\widetilde{W}$ exceeds b and $\nabla\widetilde{W}$ is equal to 0.

The intensity of the point process of locations where $\widetilde{W}$ exceeds b and $\nabla\widetilde{W}$ is equal to 0 is now considered. Let t be a fixed but arbitrary point in A and let R be a small region containing t. Denote the volume of R by vol(R). At t, the intensity of the point process of locations where $\widetilde{W}$ exceeds b and $\nabla\widetilde{W}$ is equal to 0 is

$$\lim_{\mathrm{vol}(R)\to 0} P\left\{\exists s \in R \text{ s.t. } \widetilde{W}_s \geq b\,; \nabla\widetilde{W}_s = 0\right\}/\mathrm{vol}(R)$$

The Taylor expansions

$$\widetilde{W}_t \approx \widetilde{W}_s + \nabla\widetilde{W}_s(t-s)$$

and

$$\nabla\widetilde{W}_s \approx \nabla\widetilde{W}_t + D^2\widetilde{W}_t(s-t)$$

suggest approximating the probability above by

$$\begin{aligned} & P\{\widetilde{W}_t > b\,;\ \exists s \in R \ \text{ s.t. }\ \nabla\widetilde{W}_t + D^2\widetilde{W}_t(s-t) = 0\}/\mathrm{vol}(R) \\ &\quad = P\{\widetilde{W}_t > b\,; [-D^2\widetilde{W}_t]^{-1}\nabla\widetilde{W}_t + t \ \in \ R\}/\mathrm{vol}(R) \\ &\quad = \int_b^\infty \varphi(\widetilde{w}_t)P\{[-D^2\widetilde{W}_t]^{-1}\nabla\widetilde{W}_t + t \in R \mid \widetilde{W}_t = \widetilde{w}_t\}d\widetilde{w}_t/\mathrm{vol}(R) \end{aligned}$$

To evaluate the conditional probability in the integrand above, the joint conditional distribution of $D^2\widetilde{W}_t$ and $\nabla\widetilde{W}_t$ given $\widetilde{W}_t$ is required. The content of the following computation is the idea, developed in $J7$ of Aldous (1989) and in Leadbetter et al. (1983), that conditionally, given a large value of the field at a point t, and given also that the gradient of the field at t is zero, the Jacobian of the field at t may be treated as as the constant, $-\widetilde{w}_t\Lambda_t$, and also, that the covariance of $\nabla\widetilde{W}_t$ is Λ_t. By differentiating the expressions $E\widetilde{W}_t\widetilde{W}_t = 1$

and $E\widetilde{W}_s\widetilde{W}_t = \sigma(s,t)$, and interchanging the order of differentiation and expectation, the following relations may be derived:

$$E\nabla\widetilde{W}_t\nabla\widetilde{W}_t^T = \Lambda_t;$$

$$E\nabla\widetilde{W}_t\widetilde{W}_t = 0;$$

and

$$ED^2\widetilde{W}_t\widetilde{W}_t = -\Lambda_t.$$

These relations, together with the joint normality of $\widetilde{W}$, $\nabla\widetilde{W}$ and $D^2\widetilde{W}$ suggest approximating the probability in the integrand, as vol(R) tends to 0, by

$$P\left\{[\tilde{w}_t\Lambda_t + O_p(1)]^{-1}\,\nabla\widetilde{W}_t + t \in R\right\} \sim \text{vol}(R)(2\pi)^{-\frac{d}{2}}(\tilde{w}_t^d + O(\tilde{w}_t^{d-2}))\,|\Lambda|^{\frac{1}{2}}.$$

Combining these results with the relation

$$\int_b^\infty \varphi(w)w^d dw = b^{d-1}\varphi(b) + O(b^{d-3}\varphi(b))$$

suggests approximating the intensity by

$$(2\pi)^{-\frac{d}{2}}b^{d-1}\,|\Lambda_t|^{\frac{1}{2}}.$$

Finally, consider the second step, the approximation of the probability of at least one local maximum of $\widetilde{W}$ exceeding b by the expected number of such local maxima. For the sake of simplicity, restrict attention to the interior of A, and restrict attention to points where $\widetilde{W}_t$ exceeds b and where $\nabla\widetilde{W}$ is equal to 0 rather than local maxima.

Let N_b denote the number of points where $\widetilde{W}$ exceeds b and $\nabla\widetilde{W}$ is equal to 0. In order that

$$EN_b - P\{N_b \geq 1\}$$

is of lower order than EN_b, it suffices that for all t,

$$\sup_s E\left\{N_b - 1 \mid \widetilde{W}_t \geq b\,;\ \nabla\widetilde{W}_t = 0\right\}$$

tends to 0 as b becomes large. The supremum will not tend to 0 for an arbitrary Gaussian field, but the following heuristics suggest that it should be the case for fields with the same covariance structure as W when $g(x,t)$ is chosen so that the score statistics T_t are sensitive to a localized increase in the incidence of cases.

Let t be fixed but arbitrary and consider the conditional distribution of $N_b - 1$ given that $\nabla\widetilde{W}_t$ is equal to 0 and that $\widetilde{W}_t$ is equal to $\tilde{w}_t$ for some $\tilde{w}_t$ greater than b. To observe that the conditional expectation of $N_b - 1$ is small, it is convenient to divide the points where $\nabla\widetilde{W}$ is 0 and $\widetilde{W}$ is greater than b into two subsets, those near to t and those away from t. Let R be a fixed but arbitrarily small open region containing t. The representation for $\widetilde{W}_s$ conditional on $\widetilde{W}_t = \widetilde{w}_t$ and $\nabla\widetilde{W}_t = 0$,

$$\widetilde{W}_s \approx \widetilde{w}_t(1 - \frac{1}{2}(s-t)^T\Lambda_t(s-t) + O_p((s-t)^3),$$

suggests that, if R is sufficiently small, as $\widetilde{w}_t$ becomes large, the contribution by the points in $R\backslash\{t\}$ to the conditional expectation of $N_b - 1$ tends to 0.

Let $\xi(t,s)$ denote $E\nabla\widetilde{W}_t\widetilde{W}_s$. Given $\widetilde{W}_t \geq b$ and $\nabla\widetilde{W}_t = 0$, the conditional distribution of

$$W_s^\star = \frac{\widetilde{W}_s - \sigma(s,t)\widetilde{w}_t}{\sqrt{1-\sigma(s,t)-\xi(s,t)^T\xi(s,t)}}$$

for s outside of R is that of a smooth mean 0 variance 1 Gaussian field. Let $b^\star(t,\widetilde{w}_t)$ denote

$$\sup_s \frac{b - \sigma(s,t)\widetilde{w}_t}{\sqrt{1-\sigma^2(s,t)-\xi(s,t)^T\xi(s,t)}}.$$

Then the conditional probability that $\widetilde{W}_s$ exceeds b is bounded by the conditional probability that $W_s^\star$ exceeds $b^\star(t,\widetilde{W}_t)$. It follows that if, conditionally given $\widetilde{W}_t \geq b$ and $\nabla\widetilde{W}_t = 0$, $b^\star(t,\widetilde{W}_t)$ tends to infinity, then the contribution to the conditional expectation of $N_b - 1$ by points outside of R tends to 0.

To observe that it is not unreasonable that $b^\star$ should tend to infinity for the kinds of fields considered here, note that if the $g(x,t)$ are chosen so that the score statistics are sensitive to a localized increase in the incidence of cases, then for t away from s, $g(x,t)$ and $g(x,s)$ will not be large at the same values of x. This suggests that for s outside of R, $\sigma(s,t)$ will be bounded away from 1. Also, conditionally given $\widetilde{W}_t \geq b$, $\widetilde{W}_t - b$ tends to 0 in probability as b becomes large. These two considerations, together with the definition of $b^\star$, suggest that as b becomes large, $b^\star$ will tend to infinity.

It is interesting to compare γ_b to the approximation in $J7$ of Aldous (1989) and to the approximations in Corollary 2 of Knowles and Siegmund (1989). The first term of γ_b is the natural analogue of $J7$ for non-stationary processes, and corresponds to the first term of the Knowles and Siegmund formula. The second term in γ_b corresponds to the second term of the Knowles and Siegmund formula. It would be interesting if the difference between the expected number of regions exceeding b and the probability that the field ever exceeds b could be related to the local and global overlapping discussed in remark (*ii*) following Theorem 1 of Knowles and Siegmund.

Note that when $\widetilde{W}_t$ is equal to $\widetilde{w}_t$, and a local maximum occurs at $t+h$, for h small, the approximation of $-D^2\widetilde{W}_t$ by $\widetilde{w}_t\Lambda_t$ suggests that $\widetilde{W}_{t+h}$ is approximately $\widetilde{w}_t + \frac{1}{2}h^T\Lambda_t h$. This in turn suggests that if the field is only evaluated at a finite set $t_1, t_2, ...t_M$, tail probabilities for the observed maximum might be approximated by the probability that the maximum over the whole field exceeds the envelope,

$$b(t) = b + \min_{i=1}^{M} \frac{1}{2}(t-t_i)^T\,\Lambda_{t_i}(t-t_i).$$

This in turn suggests substituting $b(t)$ for b in the formula for γ_b.

4. Examples and Simulations.

In this section, the methods are applied to the data set considered in Diggle (1991). The locations associated with the cases and controls were normalized so that the locations lay in the unit square. The function $g(x,t)$ was taken to be $e^{-\eta\|x-t\|^4}/(1+e^{-\eta\|x-t\|^4})$. with η equal to 30.0, 150.0 and 900.0. Low values of η correspond to large clusters; high values of η correspond to small clusters. For each value of η, the scan statistic was evaluated at the points $(\frac{i}{15}, \frac{j}{15})$ for i and j equal to 0, 1, ..., 15.

In order to compute γ_b, the integral in the first term,

$$\int_A |\Lambda_t|^{\frac{1}{2}}\,dt,$$

was approximated by

$$\sum_{i=1}^{15}\sum_{j=1}^{15}\left|\Lambda_{(\frac{i}{15},\frac{j}{15})}\right|^{\frac{1}{2}}/16^2.$$

That ∂A is not smooth was ignored in estimating the integral in the second term of γ_b,

$$\int_{\partial A} |\Lambda_t'|^{\frac{1}{2}}\,dt,$$

by

$$\sum_{i=0}^{15}\left\{(1,0)\Lambda_{(\frac{i}{15},0)}(1,0)^T\right\}^{\frac{1}{2}}/16 + \sum_{i=0}^{15}\left\{(1,0)\Lambda_{(\frac{i}{15},1)}(1,0)^T\right\}^{\frac{1}{2}}/16$$
$$+\sum_{j=0}^{15}\left\{(0,1)\Lambda_{(0,\frac{j}{15})}(0,1)^T\right\}^{\frac{1}{2}}/16 + \sum_{j=0}^{15}\left\{(0,1)\Lambda_{(1,\frac{j}{15})}(0,1)^T\right\}^{\frac{1}{2}}/16.$$

Figure 1 shows a plot of the case-control data treated in Diggle (1990). The cases, plotted with dark dots, are 58 cases of cancers of the larynx. The controls, plotted with light dots, are 978 cases of cancers of the lung. The

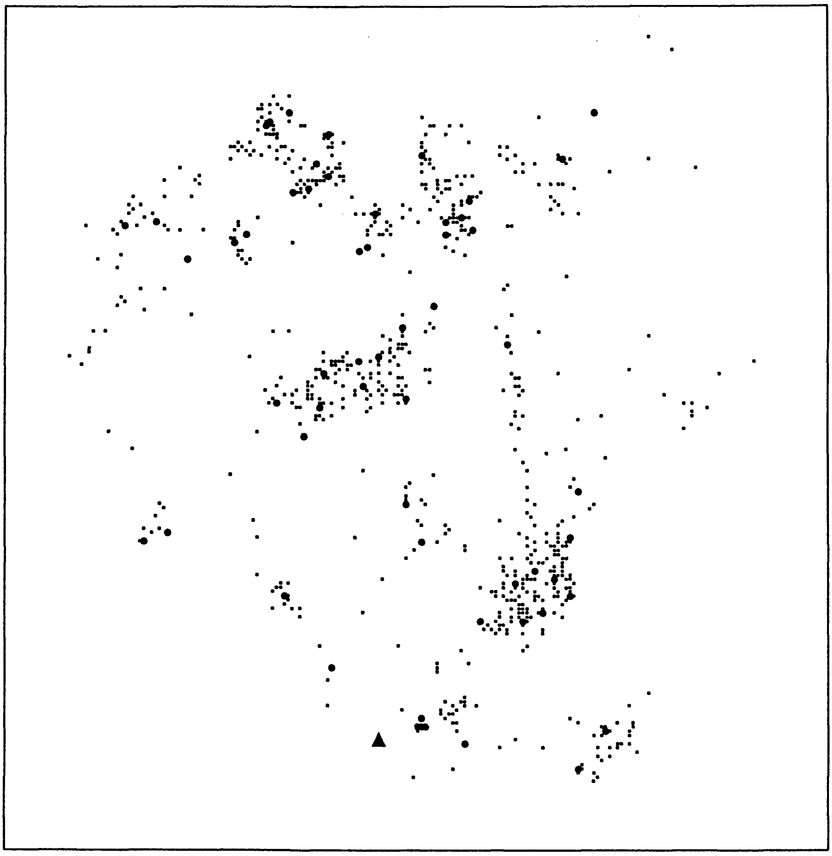

Figure 1

goal of the original analysis of this data was to determine if proximity to an incinerator appeared to be associated with an increased risk of cancer of the larynx. Proximity to the incinerator was not thought to increase the risk of cancers of the lung. Figures 2a–2c show a contour plot of the realization of the random field, for each of the three values of η, superimposed over the case-control data. The location of the incinerator is represented by a triangle.

For $\eta = 30.0$, the contour plot attains a maximum in the lower left hand corner, indicating an excess of cases on the periphery of the region where the cancers occur. When η is 150, a peak in W materializes near the incinerator. The peak is more pronounced when η is 900.0. The increased incidence of cases near the incinerator was found to be statistically significant by the methods in Diggle (1990).

Before considering the outcome of applying the approximation, γ_b, to the observed maxima in the data sets, it is useful to examine the results of some simulation experiments. In order to check the validity of the approximation in the examples, the conditional distribution under H_0 of W, and the distribution

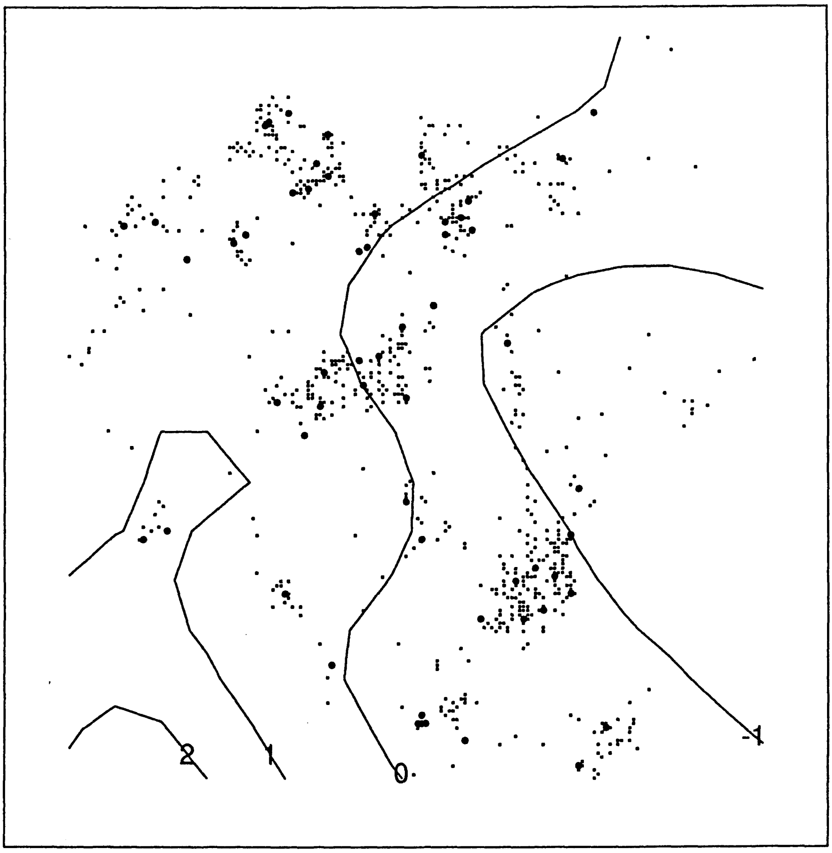

Figure 2a. Bandwidth = 30.0

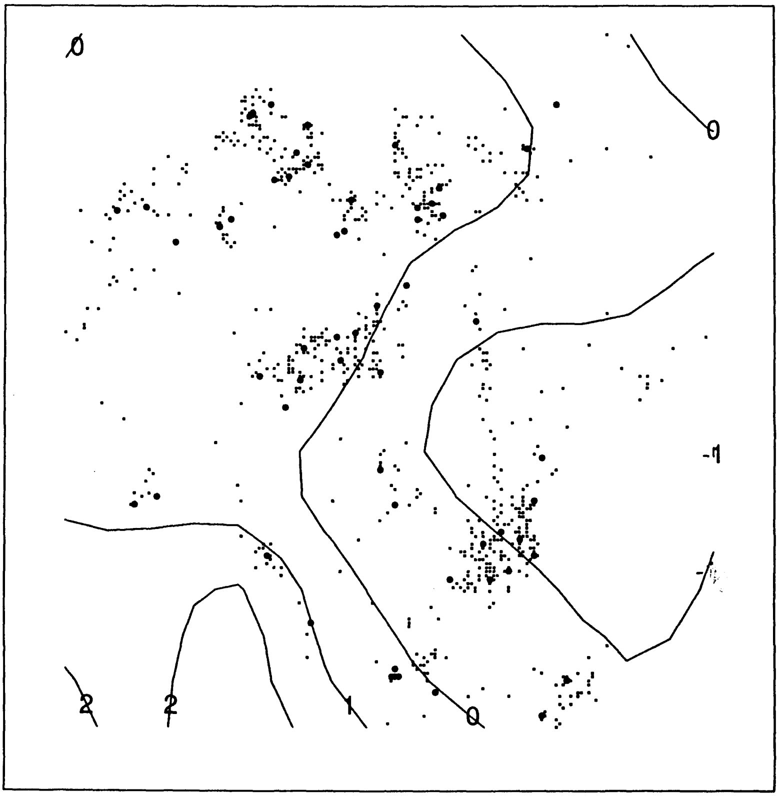

Figure 2b. Bandwidth = 150.0

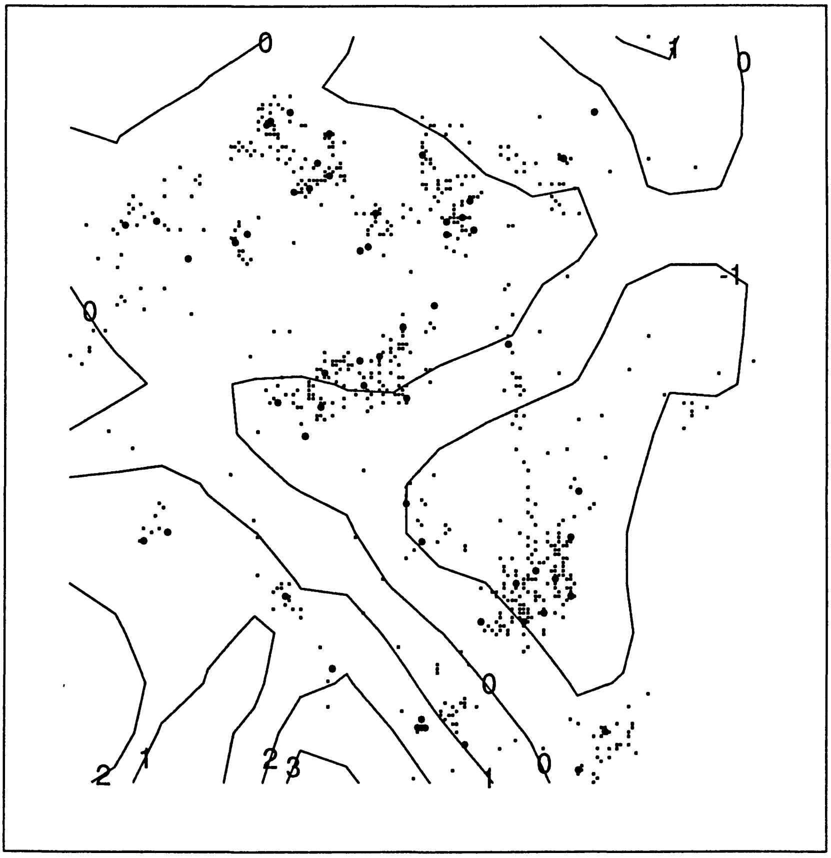

Figure 2c. Bandwidth = 900.0

of the Gaussian field $\widetilde{W}$ with the same covariance as the conditional covariance of W was simulated for each of the data sets for η equal to 30.0, 150.0 and 900.0. The simulations were based on the IMSL routines RNPER and DRNUN.

In each simulation, 900 replication of the fields were computed. With each replication, the maximum over the 16 × 16 grid was recorded. Quantiles of the empirical distributions from the simulations, together with the quantiles given by inverting the approximation γ_b are recorded in Tables 1.

It may be observed from the tables that the approximation to the quantiles given by inverting γ_b improves as the quantiles increase. Also, the approximation is a fairly accurate approximation to the distribution of the maximum of the Gaussian field. For large η, when $g(x,t)$ decreases quickly as x moves away from t, the Gaussian approximation to W is fairly inaccurate. Generally, the analytical approximation is a vast improvement over the naive p-value given by inverting the marginal $N(0,1)$ distribution, but it can be quite inaccurate even for fairly large values of b when the Gaussian approximation is poor.

Table 1
Cancer Data
Quantiles of the approximation and of the empirical distributions of the supremum of $\widetilde{W}$ and the supremum of W

	$\eta = 30.0$			$\eta = 150.0$			$\eta = 900.0$		
	αth quantile of:			αth quantile of:			αth quantile of:		
α	W	$\widetilde{W}$	approx	W	$\widetilde{W}$	approx	W	$\widetilde{W}$	approx
0.990	3.25	3.35	3.25	3.83	3.39	3.45	4.96	3.56	3.73
0.975	3.04	2.99	2.96	3.54	3.25	3.18	4.56	3.35	3.47
0.950	2.83	2.85	2.70	3.25	3.02	2.93	4.20	3.20	3.24
0.925	2.68	2.60	2.52	3.04	2.85	2.77	3.89	3.08	3.10
0.900	2.60	2.52	2.39	2.89	2.74	2.66	3.77	2.93	2.99
0.850	2.43	2.37	2.18	2.75	2.52	2.49	3.50	2.77	2.83
0.800	2.29	2.24	2.00	2.58	2.43	2.33	3.37	2.66	2.72

When the data is sparse or when the bandwidth is small, the scan statistic may depend for the most part on only a few x_i. In such cases, the Gaussian approximation to the marginal density of the field is poor, especially in the extreme tails. Substituting saddlepoint approximations to the density of the scan statistic for φ in γ_b, as in Loader (1990), might yield more accurate approximations.

Now, return to the the approximation γ_b with b equal to the maximum of W observed in the data sets. The approximation may be viewed as a nominal p-value for the null hypothesis of no localized excess risk. In the cancer data, the suprema associated with η equal to 30.0, 150.0 and 900.0 are 1.99, 2.74 and 2.94. (That there is a contour line labeled 3 in Figure 1c, is an artifact of the interpolation program that produced the plot.) The approximations to the p-value are 0.187, 0.099, and 0.123. The simulations suggest that these values are too low. The empirical probabilities from the simulations of the fields calculated with η equal to 150.0 and 900.0, are 0.193 and 0.340.

Acknowledgement. The author thanks Professor Peter Diggle for kindly supplying the cancer data set.

REFERENCES

ALDOUS, D. (1989). *Probability approximations via the Poisson clumping heuristic.* Springer-Verlag, New York.

BESAG, J. and NEWELL, J. (1991). The detection of clusters of rare diseases. *J. Roy. Statist. Soc. Ser. A* **154**, 143–155.

BERMAN, M. and EAGLESON, G. (1983). A Poisson limit theorem for incomplete symmetric statistics. *J. Appl. Probab.* **20**, 47–60.

BICKEL, P. and WICHURA, M. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42**, 1656–1670.

CUZICK, J. and EDWARDS, R. (1990). Spatial clustering for inhomogeneous populations. *J. Roy. Statist. Soc. Ser. B* **52**, 73–104.

DAY, R., WARE, J., WARTENBERG, D. and ZELEN, M. (1988). An investigation of a reported cancer cluster in Randolph, Massachusetts. *J. Clin. Epidemiol.* **42**, 137–150.

DIGGLE, P. (1990). A point process modeling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. *J. Roy. Statist. Soc. Ser. A* **153**, 349–362.

EDERER, F., MYERS, M. and MANTEL, N. (1964). A statistical problem in space and yime: Do leukemia cases come in clusters? *Biometrics* **20**, 626–638.

FRASER, D. (1983). Clustering of disease in population units: An exact test and its asymptotic version. *Amer. J. Epidemiol.* **118**, 732–739.

GATES, D. and WESTSCOTT, M. (1985). Accurate and asymptotic results for distributions of scan statistics. *J. Appl. Probab.* **22**, 531–542.

GLAZ, J. (1989). Approximations and bounds for the distribution of the scan statistic. *J. Amer. Statist. Assoc.* **84**, 560–566.

HOLST, L. (1979). Two conditional limit theorems with applications. *Ann. Statist.* **7**, 551–557.

HUNTINGTON, R. and NAUS, J. (1975). A simpler expression for kth nearest neighbor coincidence probabilities. *Ann. Probab.* **3**, 894–896.

JOHNSTONE, I. and SIEGMUND, D. (1989). On Hotelling's formula for the volume of tubes and Naiman's inequality, *Ann. Statist.* **17**, 184–194.

KNOWLES, D. and SIEGMUND, D. (1989). On Hotelling's approach to testing for a nonlinear parameter in regression. *Int. Statist. Rev.* **57**, 205–220.

LEADBETTER, M. R., LINDGREN, G. and ROOTZEN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer: Berlin, NY.

LOADER, C. (1991). Large deviation approximations to the distribution of scan statistics. *Adv. Appl. Probab.* (To appear.)

LOADER, C. (1990). Testing for nonlinearity in exponential family regression. *Bell Laboratories Technical Report*

NAGELERKE, N., OSKAM, S., FIDLER, V. and YZERMANS, C. (1988). The statistical analysis of clustering of events. *Biometrical J.* **30**, 945–956.

OHNO, Y., AOKI, K. and AOKI, N. (1979). A Test of Significance for geographic clusters of disease. *Int. J. Epidemiol.* **8**, 273–280.

RAUBERTAS, R. (1988). Spatial and temporal analysis of disease occurrence for detection of clustering. *Biometrics* **44**, 1121–1129.

ROSÉN, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement I and II. *Ann. Math. Statist.* **43** 367–397, 748–776.

SUN, J. (1990). P-values in projection pursuit. Stanford Technical Report.

STROUP, D., WILLIAMSON, D. and HERNDON, J. (1988). Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statist. Med.* **8**, 325–329.

TANGO, T. (1984). The detection of disease clustering in time. *Biometrics* **40**, 15–26.

WALLENSTEIN, S. (1980). A test for detection of clustering over time. *Amer. J. Epidemiol.* **111**, 367–372.

WEINSTOCK, M. (1981). A generalized scan statistic test for the detection of clusters. *Int. J. Epidemiol.* **10**, 289–293.

WHITTEMORE, A., FRIEND, N., BROWN, B. and HOLLY, E. (1987). A test to detect clusters of disease. *Biometrika* **74**, 631–635.

DEPARTMENT OF BIOSTATISTICS
HARVARD SCHOOL OF PUBLIC HEALTH
677 HUNTINGTON AVENUE
CAMBRIDGE, MA 02138