An isotonic spatial scan statistic for geographical disease surveillance

Martin KULLDORFF

The spatial scan statistic is commonly used in geographical disease surveillance as a test for spatial randomness to detect and evaluate geographical disease clusters. The author proposes an isotonic version of the test, where the cluster under the alternative hypothesis is modeled using an isotonic regression function with successively decreasing risk with increasing distance from the cluster center. The two methods are compared on the same breast cancer mortality data set. The result is basically the same for the two methods, detecting the same cluster, but there are subtle differences between the tests that makes one or the other the prefered choice depending on the particular situation in which they are used.

Key words: Isotonic regression, disease cluster alarms, cluster detection, spatial statistics, geomedicine

1 Introduction

Epidemiologists are often called on to investigate geographic disease cluster alarms. These may have been detected by an observant citizen group, by medical doctors, by journalists doing investigative reporting, or by local health officials. For natural reasons, the alarms are typically accompanied by considerable worry within the communities affected. Examples of past cluster alarms include: leukemia around Sellafield, England[1]; brain cancer in Los Alamos, USA[2]; leukemia in Krümmel, Germany[3]; and myeloma in Tokushima, Japan[4]. In the United States alone, the state health departments receive a total of about 1500 cancer cluster inquiries every year[5].

Most disease clusters are due to random geographical variation in disease incidence, prevalence or mortality, as some areas are bound to have higher disease rates than other, simply due to chance. On the other hand, the detection of disease clusters has sometimes led to important new knowledge, such as the discovery of serious health threats in unexpected locations[6,7], the discovery of new disease etiologies[8,9], and more rarely, the discovery of new diseases[10,11]. When disease cluster are found, it is important to use statistical methods to evaluate whether an observed excess may reasonably be due to chance or not. While not the only criteria, that will help determine whether a thorough epidemiological investigation is warranted, or whether resources are better spent elsewhere.

An alternative to post-alarm cluster analyses due to ad-hoc cluster alarms, is a pro-active approach, systematically screening a region for geographical clusters in a surveillance setting. If an ad-hoc cluster alarm subsequently does occur, it can then be quickly evaluated by looking up the result from the previous analysis, already conducted through the systematic surveillance. If there is not a significant cluster at the location in question, the alarm can be quickly dismissed as a probable chance occurrence, although some further investigation may still be warranted depending on the exact nature of the alarm. If there is a significant cluster at the location of the alarm, then the epidemiologist are not taken by surprise, but will have had a head start on investigating the disease cluster. Moreover, a systematic approach to geographical disease surveillance may detect important disease clusters that would otherwise go unnoticed.

One method that can be used for geographical disease surveillance as well as for the evaluation of geographical disease cluster alarms, is the spatial scan statistic[12,13]. The spatial scan statistic has the following features, making it suitable for geographical disease surveillance: (1) it takes the uneven geographical distribution of the population risk into account, and adjusts for any number of confounding variables, such as age, gender or other known or suspected risk factors; (2) it searches for clusters without making any apriori

Division of Biostatistics
Department of Community Medicine and Health Care
University of Connecticut School of Medicine
Farmington, CT 06030-6205, United States

assumptions about their location or size; (3) it adjusts for the multiple testing inherent in the multiple cluster locations and sizes considered; (4) if the null hypothesis is rejected, it specifies the location of the cluster that caused the rejection; (5) it is able to detect and evaluate multiple clusters if they exist; and (6) it can be used to detect areas of excess risk (clusters) as well as areas of lower risk.

In this paper, we first describe the spatial scan statistic. We then present an isotonic spatial scan statistic. It has the same properties as listed above for the standard spatial scan statistic. Both are constructed as a likelihood ratio tests, but they differ in the way in which clusters are modeled under the alternative hypothesis. The isotonic spatial scan statistic is illustrated using a breast cancer mortality data set previously analyzed using the standard spatial scan statistic[14], and the two methods are compared.

2 The Spatial Scan Statistic

The spatial scan statistic imposes a circular window on a map and lets its centroid move across the study region. For any given position of the centroid, the radius of the window is changed continuously to take any value between zero and some upper limit. In total the method uses a set Z containing a very large number of distinct circles, each with a different location and size, and each being a potential cluster. A small sample of the many circles used are shown in figure 1.

Under the alternative hypothesis, there is at least one circle for which the underlying risk is higher inside the circle as compared to outside. For each circle, the observed and expected number of cases are noted. It is then possible to calculate the likelihood to observe the actually observed number of cases within a circle. The circle with the maximum likelihood is defined as the *the most likely cluster*. This is the cluster that is least likely to have occured by chance.

The likelihood can be calculated assuming either a Poisson or Bernoulli model, depending on how the cases are generated. The Poisson model should be used for incidence and mortality rates, where the population at risk reflect the number of person years. The Bernoulli model should be used if we have 0/1 observations such as children with birth defects compared to total births, or late stage breast cancer as compared to the total cases of breast cancer.

Conditioning on the observed total number of cases, N, the definition of the spatial scan statistic is the maximum likelihood ratio over all possible circles $Z \in \mathcal{Z}$

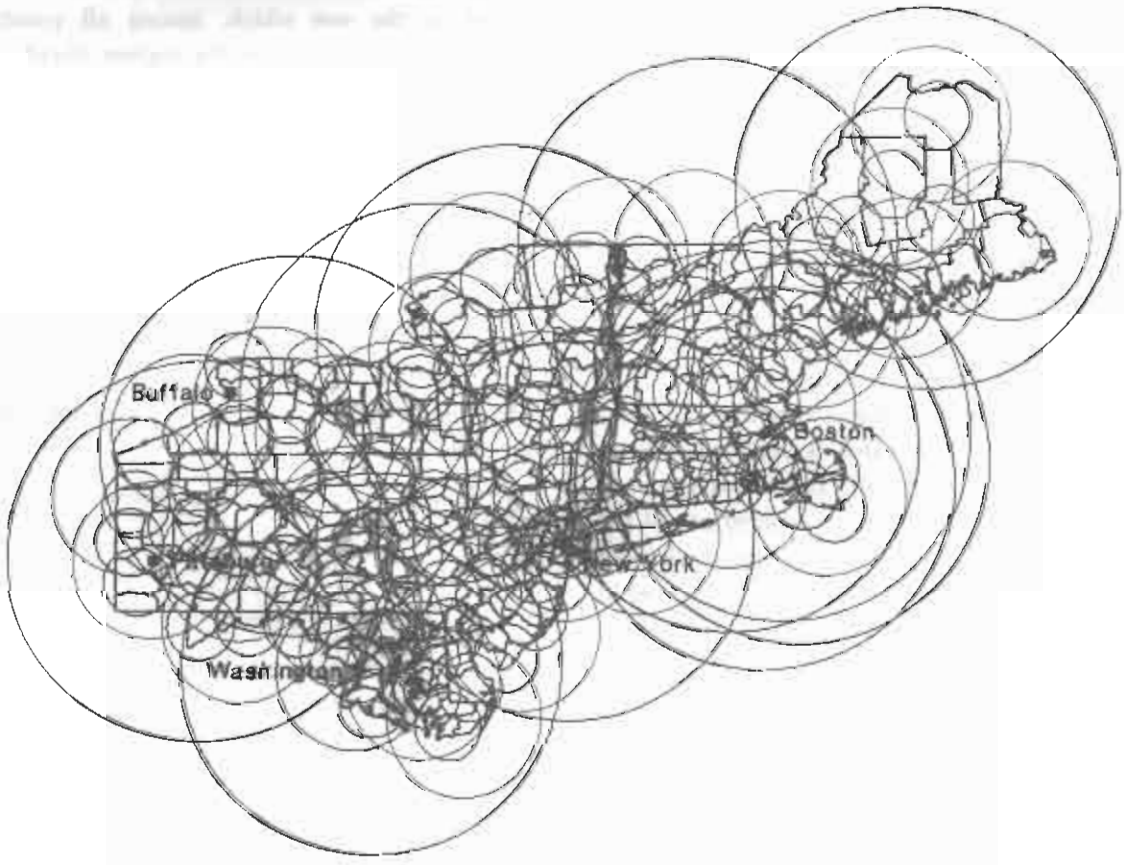

Figure 1 A small sample of the many circles used by the spatial scan statistic

$$S_z = \frac{\max_{Z \ni z} L(Z)}{L_0} = \max_{Z \ni z} \frac{L(Z)}{L_0} \quad (1)$$

where $L(Z)$ is the maximum likelihood for circle Z, expressing how likely the observed data are given a differential rate of events within and outside the zone, and where L_0 is the likelihood function under the null hypothesis.

Let n_Z be the number of cases in circle Z. For the Bernoulli model, let M be the total number of cases and controls, and let m_Z be the combined number of cases and controls in circle Z. Then

$$L(Z,p,q) = p^{n_Z}(1-p)^{m_Z-n_Z} q^{N-n_Z}(1-q)^{M-N-m_Z+n_Z} \quad (2)$$

where p is the probability that an individual within zone Z is a case and where q is the same probability for an individual outside the zone. Maximizing the likelihood over p and q gives

$$\frac{L(Z)}{L_0} = \frac{\max_{p>q} L(Z,p,q)}{\max_{p=q} L(Z,p,q)} = \frac{\left(\frac{n_Z}{m_Z}\right)^{n_Z}\left(1-\frac{n_Z}{m_Z}\right)^{m_Z-n_Z}\left(\frac{N-n_Z}{M-m_Z}\right)^{N-n_Z}\left(1-\frac{N-n_Z}{M-m_Z}\right)^{M-N-m_Z-n_Z}}{\left(\frac{N}{M}\right)^{N}\left(1-\frac{N}{M}\right)^{M-N}} \quad (3)$$

if $n_Z/m_Z > (N-n_Z)/(M-m_Z)$, and one otherwise.

For the Poisson model, let $\mu(Z)$ be the expected number under the null hypothesis, so that $\mu(A)=N$ for A, the total region under study. It can then be shown that

$$\frac{L(Z)}{L_0} = \left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{N-n_Z}{N-\mu(Z)}\right)^{N-n_Z} \quad (4)$$

if $n_Z > \mu(Z)$ and one otherwise. Details, including derivations as likelihood ratio tests, have been given elsewhere[12].

As this likelihood ratio is maximized over all the circles, it identifies the one that constitutes the most likely disease cluster. Its p value is obtained through Monte Carlo hypothesis testing[13]. Conditioning on the total number of cases, 9999 random replications of the data set are generated under the null hypothesis. For each of these, the maximum likelihood is calculated in the same way as for the real data set. The 10,000 values, from the real and random data sets, are then ranked from highest to lowest. If the null hypothesis is true, then the maximum likelihood ratio from the real data set has the same distribution as the maximum likelihood ratio calculated from any of the random data sets. Hence, its rank R on the list is uniformly distrubuted between 1 and 10,000, that is, it is equally likely to take any of the 10,000 positions among the ranked values. This means that the probability of being among the top 5 percent values is, exactly, 5 percent, and its p value can be calculated as $p = R/10000$.

Calculations can be done using the SaTScan software[16] developed at the National Cancer Institute, and available free of charge.

3 An Isotonic Spatial Scan Statistic

For the standard spatial scan statistic, and for a given centroid, the risk is modeled as being higher within some unknown distance d from the centroid, as compared to beyond that distance. The distance d correspond to the radius of the circle used, and as mentioned before, this radius is not specified apriori. This means that the risk is modeled as a function $r(d)$ of the distance from the centroid, and that it uses a step function with a single discontinuity at d.

Such a formulation leads to a natural extention of the spatial scan statistic, in that the risk function could be modeled as a non-increasing function with multiple locations where the function takes a step down, little by little. If we don't make any apriori assumptions about the number or locations of those steps, the risk function can be fitted using maximum likelihood in what is called *isotonic regression*[17]. The isotonic regression function is defined as the one which, among all possible non-increasing functions, gives the highest likelihood.

Mathematically, the only difference as compared to the standard version is $L(Z)$. Firstly, rather than using a large set of circles, we will use a large set $\boldsymbol{C}$ of circle centroids. For a particular centroid $C \in \boldsymbol{C}$, order the census areas in order of distance from the centroid, with the closest first. Let n_j, $j=1, \ldots, J$ be the number of cases in census area j.

For the Bernoulli model, let p_j, $j=1, \ldots, J$ be the probability that an individual in census area j is a case, and let m_j be the total number of cases and controls in that area. Let

$$L(C, p_1, \ldots, p_J) = \prod_{j=1}^{J} p_j^{n_j}(1-p_j)^{m_j-n_j} \quad (5)$$

We then define

$$L(C) = \max_{p_1 \geq p_2 \geq \cdots \geq p_J} L(C, p_1, \ldots, p_J) \quad (6)$$

and

$$L_0 = \max_{p_1 = p_2 = \cdots = p_J} L(C, p_1, \ldots, p_J) \quad (7)$$

The test statistic is

$$S_C = \max_{C \in \boldsymbol{C}} \frac{L(C)}{L_0} \quad (8)$$

Note that L_0 does not depend on C, and has the same value as for the standard spatial scan statistic.

For the Poisson model, let μ_i be the expected number of cases in census area i (under the null hypothesis). The test statistic is then

$$S_C = \max_{z,\lambda_1,\lambda_2,\ldots,\lambda_J} \frac{\prod_{j=1}^{J}(\lambda_j \mu_j)^{w_j}}{\left(\sum_{j=1}^{J} \lambda_j \mu_j\right)^{N}} \qquad (9)$$

where N as before is the total number of cases.

The collection of centroids C can be chosen in different ways, but for the test statistic to be a scan statistic, searching the area for clusters, it is important that the centroids are at least as close to each other as are the census areas, and that they cover the whole geographical area under study. In the forthcoming application, described in the next section, we simply used the coordinates of the 245 counties as circle centroids.

If there is only one centroid in C, we do no longer have a scan statistic, but rather what is called a focused cluster test[18,19,20,21,22]. In fact, we get the T_1 test proposed by Stone for focused cluster analyses, although he did not implement it due to the difficulty in obtaining an asymptotic distribution of the test statistic, using only the first isotonic regression estimator instead[20]. Focused tests are used to see whether a disease cluster exist around some potential putative health hazard at some specified foci, when that foci is specified apriori, without first looking at the case counts. For example, Lawson has used such a test to see whether there were a cluster of respiratory cancer around a source of air pollution in Armadale, Scotland[23].

If a suspected health hazard exist before looking at the data, a focused test should be used rather than a scan statistic, as the focused test will have higher power to detect a cluster in that particular location. If on the other hand, a cluster is first found, a focused test cannot be used, as they would be erroneous p-values due to preselection bias. To evaluate cluster alarms, one should instead use the spatial scan statistic, as it does not make apriori assumptions on the cluster location and size.

4 Breast Cancer in Northeastern United States

Kulldorff et al. used the standard spatial scan statistic to study the geographical distribution of female breast cancer mortality in Northeastern United States[14]. This data set encompasses the years 1988–1992 and covers the 245 counties and county equivalents in Connecticut, Delaware, District of Columbia, Maine, Maryland, New Hampshire, New Jersey, New York, Pennsylvania,

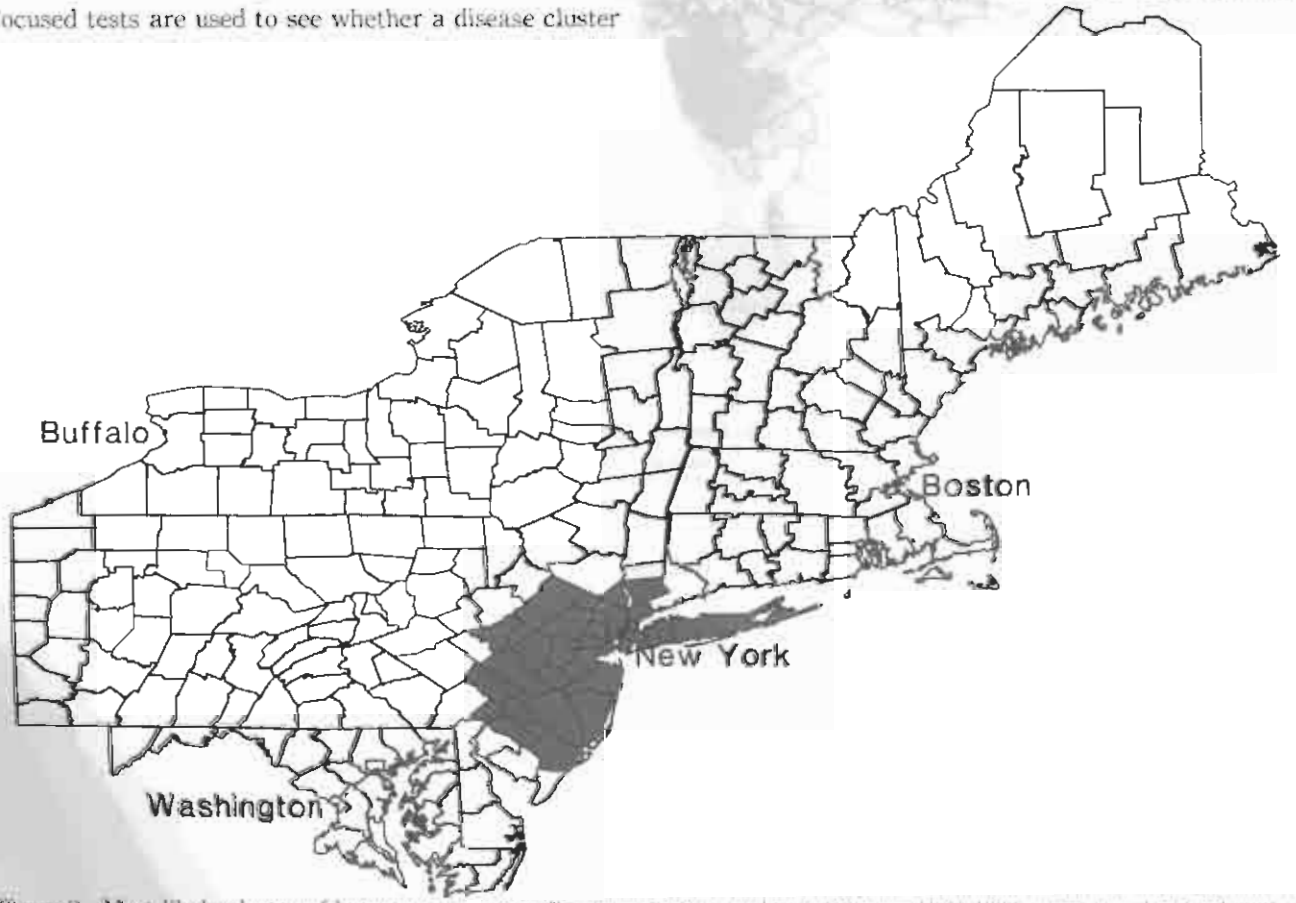

Figure 2 Most likely cluster of breast cancer mortality in northeastern United States, 1988–1992, using the standard spatial scan statistics

Table 1 Most likely cluster of breast cancer mortality in the Northeastern United States, 1988–1992, comparing the standard versus the isotonic spatial scan statistic.

Test	Most Likely Cluster	Counties	Cases	Expected	RR†	LLR‡	p value
Standard	NYC-Philadelphia Area	32	24044	23040	1.074	35.7	0.0001
Isotonic	NYC-Philadelphia Area	40	25582	24601	1.070	42.7	0.0001

†Relative risk compared to the rest of the Northeast, outside the cluster.
‡Log likelihood ratio.

Figure 3 Most likely cluster of breast cancer mortality in northeastern United States, 1988–1992, using the isotonic spatial scan statistics.

Table 2 Most likely cluster of breast cancer mortality in the Northeastern United States, 1988–1992, using the isotonic spatial scan statistic.

Step	Counties	Cases	Expected	RR†
1	Ocean, Burlington, Monmouth Atlantic, Mercer, Camden Middlesex, Philadelphia	5460	4991	1.126
2	Gloucester	215	198	1.115
3	Richmond, Somerset, Union	1183	1134	1.074
4	Remaining 28 counties	18724	18278	1.055
Combined		25582	24601	1.070

†Relative risk compared to the rest of the Northeast, outside the cluster.

Rhode Island, and Vermont. There were a total of 58,943 cases among a population of 29,535,210 women. The annual mortality rate was 39.9 per 100,000 women. All analyses in this paper uses the Poisson model, and are adjusted for age using 18 different five year age groups: 0-4, 5-9, ..., 80-84, and 85+.

Figure 2 and table 1 show the result using the standard spatial scan statistic. The most likely cluster was centered around Monmouth county in New Jersey, just south of New York City along the Atlantic ocean. The cluster contains most of the New York City-Philadelphia metropolitan area. With a total of 24,044 cases compared to 23,040 expected, the mortality rate is 7.4 percent higher than the remaining parts of Northeastern United States. The cluster is significant with a p-value of 0.0001. A secondary cluster was found around Buffalo, but this cluster was not statistically significant ($p=0.12$).

The result for the isotonic spatial scan statistic is shown in table 1 and figure 3. The most likely cluster is now centered around Ocean county, just south of the previous center in Monmouth county. The general location and size is the same as for the standard spatial scan statistic, but somewhat expanded southward to include eight additional counties.

The isotonic step function has four levels, as shown in table 2. There is a fairly large inner circle, with an excess risk of 12.6 percent. There is then two smaller steps, with only one and three counties in them respectively, with successively lower risk. The fourth step is again large, containing 28 counties, but the excess risk is now only 5.5 percent. The cluster as a whole has a 7.0 percent excess risk, as compared to the 7.4 percent excess found in the somewhat smaller cluster from the standard spatial scan statistic.

The isotonic spatial scan statistic also detects a secondary cluster in the Buffalo area, with a two step isotonic risk function. There are three counties in the inner circle (Cattaraugus, McKean, Allegany) and one in the outer (Erie). It is not statistically significant though, with $p=0.34$. Other non-significant secondary clusters contained anywhere between 2 and 6 steps in the isotonic regression function.

5 Discussion

For the breast cancer mortality data, the result is basically the same whether one uses the standard or the isotonic spatial scan statistic. As has been pointed out previously[14], the spatial scan statistic can detect the general location and size of a cluster, but its exact boundaries must remain uncertain. This is because the likelihood will not change much when adding or removing a few smaller counties to or from the most likely cluster. The same observation is valid for the isotonic

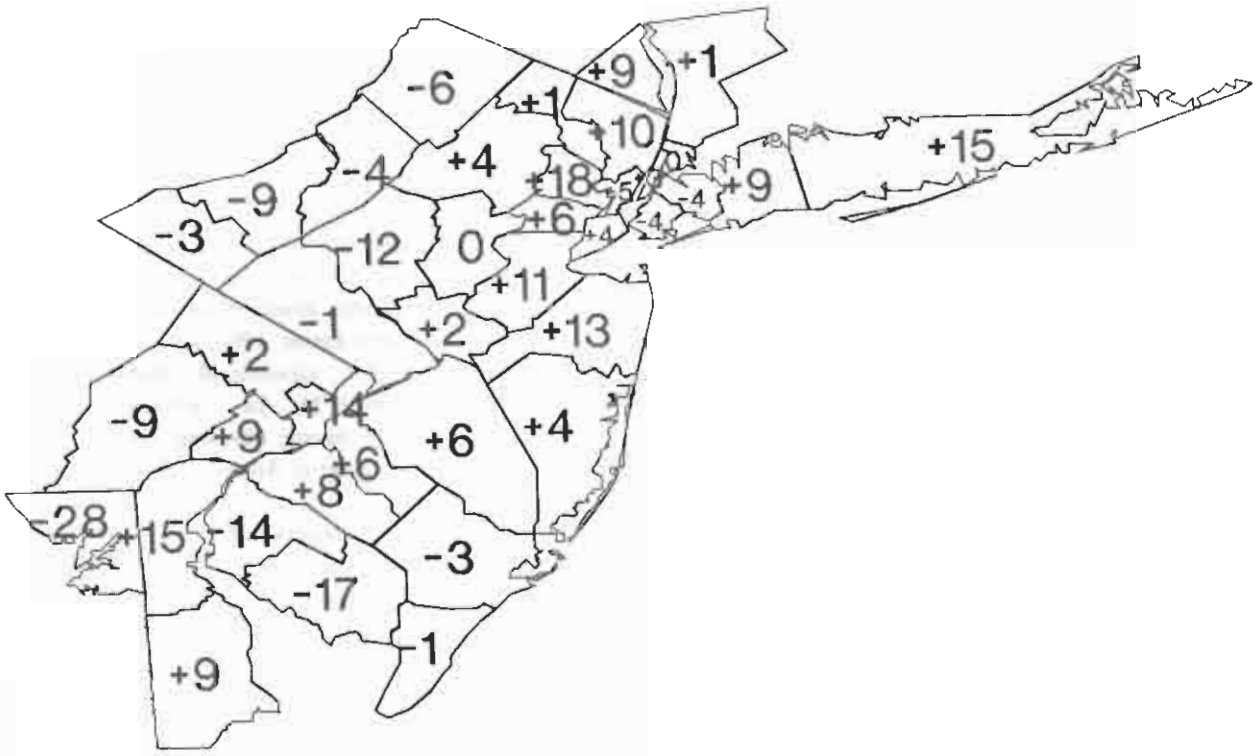

Figure 4 County specific mortality rates: Percent above or below expected.

spatial scan statistic, and not only for the cluster as a whole, but also for the intermediate steps in the risk function.

So, which method should be used, the standard or the isotonic spatial scan statistic? How do they differ? What are their pros and cons? We cannot compare the methods by the magnitude of their respective log likelihood ratio values. Since the circles in the standard version are all special cases of the isotonic risk model, it is mathematically impossible for the standard version to have a higher log likelihood ratio than the isotonic version.

One apparent advantage of the isotonic version is that we obtain information about which parts of the cluster has higher versus lower rates. That information can be deceptive though, as figure 4 illustrate. There are several counties in the outer band that has an excess as high or higher than the central counties, while one county in the inner band has fewer cases than expected. Of course, these individual county estimates are highly variable, but it illustrates the point already raised, that while we can pinpoint the general location and size of a cluster, we can say nothing about its exact boundaries, and even less about the variability in risk within the cluster. Hence, the variable rates within the cluster, as produced by the isotonic spatial scan statistic, may be more likely to confuse the reader of the map than to provide useful information. For descriptive details, it is better to provide the type of map given in Figure 4.

One difference between the two tests concerns statistical power. The isotonic version will have somewhat higher power if the true cluster is such that the risk is higher in the center of the cluster as compared to the edges of the cluster, but it will have somewhat lower power if the true cluster has approximately equal risk throughout the cluster area or if the risk is higher in the outlying as compared to the central areas of the cluster. The magnitude of the power difference is unknown, and merits further investigation, but is likely to be fairly small.

Some geographical phenomena are very gradual, such as air pollution or access to health care, while other display abrupt changes, such as sources of water supply and certain socioeconomic variables. The choice of method should depend on what types of exposure we think are most likely to cause any clusters that we might find, which of course in a surveillance setting, is not easy to say.

Both methods are computer intensive, in the sense that they are based on Monte Carlo replications for calculating p values. The difference is computing time is not large though. For the current data set, using 9999 Monte Carlo replications, the isotonic spatial scan statistic needed 4 minutes to run on a 400 MHz Pentium PC, while the standard spatial scan statistic took 10 minutes to run. The standard version is slower not because the computations are more complex, but because the code used is optimized for situations with many census areas. With fewer cases but more census areas, the comparison would have been the opposite, with the standard version using less computing time.

In summary, there are sometimes advantages of using the isotonic spatial scan statistic, but the simplicity of concept and interpretation will tend to favor the standard spatial scan statistic in many situations.

References

1) Gardner MJ, Snee MP, Hall AJ, Powell CA, Downes S, Terrell JD. 1990. Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria. British Medical Journal, 300: 423-429.

2) Kulldorff M, Feuer EJ, Miller BA, Athas WF, and Key CR. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. American Journal of Public Health, 1998; 88: 1377-1380.

3) Dieckmann H. Haufung von Leukamieerkrankungen in der Elbmarsch (Incidence of leukemia in the Elbmarsch area). Gesundheitswesen, 1992; 54: 592-596.

4) Kosaka M, Okagawa K, Miyamoto Y, Goto T, Saito S. Geographic clustering of myeloma in Tokushima. International Journal of Hematology, 1991; 54: 405-409.

5) Wartenberg D, Greenberg M. Characterizing cluster studies: A review of the literature. Presentation at the conference on Statistics and Computing in Disease Clustering, Vancouver, Canada, 1994.

6) Czeizel AE, Elek, Gundy S, Metneki J, Nemes E, Reis A, Sperling K, Timar L, Tusnady G, Viragh Z. Environmental trichlorfon and cluster of congenital abnormalities. Lancet, 1993; 341: 539-542.

7) Public Citizen Health Research Group. Who poisoned the children? Health Letter, 1997; 13: 12: 3-6.

8) Baris YI, Sahin AA, Ozesmin M, Kerse I, Ozen E, Kolacan B, Ogankulu M: An outbreak of pleural mesotheliomas and chronic fibrosing pleurisy in the village of Karain/Urgalp in Anatolia. Thorax, 1978; 33: 181-192.

9) Uhlig M. Über den Schneeberger Lungenkrebs (Concerning the Schneeberger lung cancer). Virchows Archiv für Patologische Anatomie, 1921; 230: 76-98.

10) Gottlieb MS, Schanker HM, Fan PT, Saxon A, Weisman JD, Pozalski I. Pneumocystis pneumonia-Los Angeles. Mortality and Morbidity Weekly Report, 1981; 30: 250-252.

11) Steere AC, Malawista SE, Snydman DR, Shope RE, Andiman WA, Ross MR, Steele FM. Lyme arthritis: an

epidemic of oligioarticular arthritis in children and adults in three Connecticut communities. Arthritis and Rheumatism. 1977; 20: 7-17.
12) Kulldorff M. A spatial scan statistic. Communications in Statistics: Theory and Methods, 1997; 26: 1481-1496.
13) Kulldorff M, Nagarwalla N. Spatial disease clusters: Detection and inference. Statistics in Medicine, 1995; 14: 799-810.
14) Kulldorff M, Feuer E, Miller B, Freedman L. Breast cancer in northeast United States: A geographic analysis. American Journal of Epidemiology, 1997; 146: 161-170.
15) Dwass M. Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics. 1957; 28: 181-187.
16) Kulldorff M, Rand K, Gherman G, Williams G, DeFrancesco D. SaTScan v 2.1: Software for the spatial and space-time scan statistics. Bethesda, MD: National Cancer Institute, 1998. *(http://dcp.nci.nih.gov/BB/SaT-Scan.html)*
17) Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD. Statistical inference under order restrictions: The theory and application of isotonic regression. Chichester, England: Wiley, 1972.
18) Bithell JF. The choice of test for detecting raised disease risk near a point source. Statistics in Medicine, 1995; 14: 2309-2322.
19) Lawson AB. On the analysis of mortality events associated with a prespecified fixed point. Journal of the Royal Statistical Society, Series A, 1993; 156: 363-377.
20) Stone RA. Investigation of excess environmental risk around putative sources: statistical problems and a proposed test. Statistics in Medicine, 1988; 7: 649-660.
21) Waller LA. Statistical power and design of focused clustering studies. Statistics in Medicine, 1995; 15: 765-782.
22) Waller LA, Lawson AB. The power of focused tests to detect disease clustering. Statistics in Medicine, 1995; 14: 2291-2308.
23) Lawson A, Williams FL. Armadale: A case study in environmental epidemiology. Journal of the Royal Statistical Society, 1994; A157: 285-298.

本論文では，疾病の地域集積性の検出のために著者が開発した Spatial Scan Statistic の一つの修正版を提案している．それは，疾病の集積性を表現する一つのモデルとして，集積の中心地(cluster center)からの距離と疾病のリスクが反比例するという isotonic 回帰関数を利用している．オリジナルの方法では，疾病集積が見られる地域は，そうでない地域に比較して，リスクが高いという単純なモデル(hot spot model)であった．アメリカ北東部の乳癌死亡データに適用して，二つの方法を比較したところ，ほとんど差がなかったようである．一般には，集積性の構造によって二つの方法の性能に違いが期待されるので，二つのモデルの違いは有用であると結論している．