Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2010, Article ID 642379, 11 pages
doi:10.1155/2010/642379

Research Article

Spatial Scan Statistics Adjusted for Multiple Clusters

Zhenkui Zhang,[1] Renato Assunção,[2] and Martin Kulldorff[3]

[1] *Personal Market - Property Strategic Research Team, Liberty Mutual Group, 175 Berkeley Street 10GH, Boston, MA 02116-4715, USA*
[2] *Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte, MG, Brazil*
[3] *Department of Population Medicine, Harvard Pilgrim Health Care Institute, Harvard Medical School, 133 Brookline Avenue, Boston, MA 02215, USA*

Correspondence should be addressed to Renato Assunção, assuncao@est.ufmg.br

Received 26 November 2009; Revised 24 March 2010; Accepted 9 June 2010

Academic Editor: Rongling Wu

The spatial scan statistic is one of the main epidemiological tools to test for the presence of disease clusters in a geographical region. While the statistical significance of the most likely cluster is correctly assessed using the model assumptions, secondary clusters tend to have conservatively high P-values. In this paper, we propose a sequential version of the spatial scan statistic to adjust for the presence of other clusters in the study region. The procedure removes the effect due to the more likely clusters on less significant clusters by sequential deletion of the previously detected clusters. Using the Northeastern United States geography and population in a simulation study, we calculated the type I error probability and the power of this sequential test under different alternative models concerning the locations and sizes of the true clusters. The results show that the type I error probability of our method is close to the nominal α level and that for secondary clusters its power is higher than the standard unadjusted scan statistic.

1. Introduction

Spatial and space-time scan statistics [1] have become some of the main tools in geographic disease surveillance to test the null hypothesis that geographical data are randomly distributed against a localized cluster alternative. Examples include its use for breast cancer mortality in Texas [2], giardiasis parasites in Canada [3], pneumonia in Brazil [4], lymphatic filariasis in Haiti [5], and syndromic surveillance in New York City [6].

The standard spatial scan statistic is a maximum likelihood ratio test statistic S based on a circular window of variable size scanning the geographical area under surveillance. Under the null hypothesis, there are no disease clusters. Under the alternative hypothesis, there is a single geographical cluster in the region of unknown location and size. A detailed description is presented in Section 2.

When the test significantly detects one cluster, it is of interest to know if there are additional clusters present in the region. The test procedure provides evidence for the presence of these so-called *secondary clusters*, spatial clusters not overlapping with the most likely cluster but with significantly large likelihood ratio. These secondary clusters have an associated P-value but they are calculated ignoring the existence of the most likely cluster that had already been detected. One consequence of this is that these are conservative P-values [1] leading to a loss in statistical power. The P-values for testing for a second cluster could alternatively be calculated conditionally on the presence of the primary cluster and their P-values could then be smaller than those delivered by the standard method.

In this paper, we propose a sequential method to evaluate the statistical significance of secondary clusters by removing the effect caused by the previously detected stronger clusters. In brief, data from the most likely cluster (MLC), possibly together with some neighboring tracts, are deleted from the dataset if the MLC is statistically significant. The standard spatial scan statistic is then applied to the reduced dataset to test for the presence of a second cluster. The procedure is reiterated until no further statistically significant cluster is found. We also evaluate an alternative approach, where we replace the cases in the most likely cluster with the expected number of cases for each location in that cluster. The expected numbers are calculated using data only from the tracts outside the MLC.

Either methods is valid only if their type I error probability is close to the nominal α level. Our simulation results show that the type I error probability is under control. We compare the power of testing the significance of the second most likely cluster between the standard method and the sequential method finding that the sequential method has higher power for secondary clusters.

The paper is organized as follows. The spatial scan statistic is briefly reviewed in Section 2. In Section 3, we adjust the spatial scan statistics by either removing the most likely cluster or by replacing the most likely cluster with expected counts. We present the results concerning the type I error probability and statistical power in Section 4. A practical example applying both the standard approach and the new sequential approach to breast cancer mortality in the Northeastern United States is given in Section 5. In Section 6, we discuss the results derived in this paper.

2. The Spatial Scan Statistic

Suppose a geographical region is partitioned into small areas such as, for example, zip code areas or census tracts. The areas are represented by their geometrical centroids. The data input for the spatial scan statistic includes the geometrical centroid location (longitude and latitude coordinates), the population at risk count, and the number of disease cases in each area. We assume that the counts are independent Poisson distributed random variables. The risk population may be the raw population count or a covariate adjusted population at risk estimate. Centered at each centroid, a collection of circles of continuously varying radii defines the potential clusters: each one is composed by its center plus the neighboring centroids contained within the circle.

Let n_z and c_Z denote the total covariate adjusted population at risk and the total number of cases of the areas within circle Z, respectively. In most disease surveillance applications, there is no reliable external estimate concerning the expected number of cases under the null hypothesis. Also, interest is most often in comparing disease risk in different parts of the map rather than a comparison with an external region or time period. Hence, we use the conditional rather than unconditional scan statistic [7]. Conditioning on the observed total number C of cases, the spatial scan statistic S is the maximum of the likelihood ratio over all possible circles Z:

$$S = \frac{\max_Z L(Z)}{L_0} = \max_Z \frac{L(Z)}{L_0}, \tag{2.1}$$

where $L(Z)$ is the likelihood that circle Z has c_z observed cases conditionally on the total number of cases C, on the within Z population n_z, and on the total population N. The denominator L_0 is the likelihood function under the null hypothesis.

If we know the null hypothesis underlying disease rate, we can approach this problem in a nonconditional way, as in [8]. He shows that better performance can be possible for the spatiotemporal scan method when one does not condition on the total number of cases. However, in most applications one does not have an estimate of the underlying rate with enough precision to assume it known.

Let $\mu(Z) = n_z C/N$ be the expected number of cases in circle Z. We have $L(Z)/L_0 \geq 1$ for all Z with $L(Z)/L_0 = 1$ if $c_Z < \mu(Z)$. Therefore, we can write

$$S = \max_Z \frac{L(Z)}{L_0} = \max_Z \left(\frac{c_Z}{\mu(Z)}\right)^{c_Z} \left(\frac{C - c_Z}{C - \mu(Z)}\right)^{C - c_Z}. \tag{2.2}$$

The α level critical value s is defined as $P(S > s \mid H_0) = \alpha$, where α is the type I error probability, and H_0 represents the null hypothesis. A circle with likelihood ratio $S > s$ is significant, and the circle Z with the maximum likelihood ratio S is called the most likely cluster (MLC). Besides the MLC, multiple additional clusters can be derived and evaluated as well. The most interesting additional clusters are the circles Z that do not overlap with the MLC and that have the likelihood ratio $L(Z)/L_0$ larger than the critical cutoff point s.

To evaluate the significance, the scan statistic test uses Monte Carlo hypothesis testing [9]. The P-value of the MLC is calculated by repeatedly simulating data under the null hypothesis conditional on the same total number of cases C. For each of M simulations one calculates the maximum likelihood ratio statistic. The critical value s is equal to the $\alpha(M+1)$ highest of these M maximum likelihood ratio statistics and $p = R/(M+1)$ where $R-1$ is the number of test statistics from the simulated datasets that are larger than the test statistic S from the real data.

If the test is significant, there is interest in testing for the presence of additional secondary clusters that do not overlap with the most likely cluster. The standard way to do this is to compare the likelihood ratio of such secondary clusters with the maximum likelihood ratios from the simulated data. The interpretation of this approach is that we are evaluating whether the secondary clusters are able to reject the null hypothesis on their own strength, whether the most likely cluster is a true cluster or not. In some sense, it is like a regression analysis, where each variable is entered in a separate regression model and evaluated without adjusting for other variables. A drawback of this approach is that the P-values are conservative [1] with a corresponding loss of statistical power. This is because the likelihood ratio from the real data is less than the maximum while it is compared to the maximum likelihood ratios from the simulated datasets.

An alternative approach would be to compare the likelihood ratios from secondary clusters in the real data with the likelihood ratios from the corresponding secondary clusters from the simulated data. However, since the simulations are carried out under the null hypothesis, these likelihood ratios from the simulations do not take into account that there is one cluster already present in the map and outside the secondary cluster. A more interesting approach is to try and adjust for the most likely clusters when evaluating secondary clusters.

3. Adjusting the Spatial Scan Statistic for Multiple Clusters

We propose to test the statistical significance of multiple clusters sequentially, so that one would test the second most likely cluster only if the MLC is significant, and the third most likely cluster only if the second most likely cluster is significant and so on. Since the spatial scan statistic calculates the likelihood of cases within the circle to the remaining cases outside the circle, the previously detected most likely clusters have an effect on calculating the test statistic for the less likely clusters. Our objective is to eliminate this effect to be able to test for the presence of additional clusters conditional on the presence of the previously detected clusters.

To reduce the effect due to the MLC, we remove from the original data the areas comprising the MLC. We also experiment with removing not only the MLC but also some of its nearest neighboring areas. The areas that are removed leave an empty region in the map, with no population and no cases. Next, the scan test procedure is carried out on this reduced dataset, treating the removed area as if it was a "lake" with no information. This procedure is iterated until no further statistically significant clusters are found.

By a conditioning argument, we can justify this deletion of the first cluster from the map to run the second stage test. Suppose that we want to test for the presence of a second cluster Z_2 given that we detected *exactly* the first cluster Z_1. Since we assume independent counts in the areas, the likelihood can be written as the product $L_1(p_1)L_2(p_2)L_r(p_r)$ of three factors. The first one, $L_1(p_1)$ is associated only with the counts in Z_1 with probability p_1. The second one is similarly defined for a second cluster Z_2 with null intersection with Z_1 and the third is associated with the remaining of the areas ($L_r(p_r)$). Then, the distribution of the counts conditioned on those observed in Z_1 is simply $L_2(p_2)L_r(p_r)$ and the maximum likelihood test statistic should be

$$\max_{Z_2 : Z_2 \cap Z_1 = \emptyset} \frac{\max_{p_2 > p_r} \max L_2 L_r}{\max_{p_2 = p_r} L_2 L_r}. \tag{3.1}$$

This is the same test statistic one obtains by removing from the map the data from cluster Z_1 and working in the reduced map with the usual scan statistic test.

An alternative approach is to replace the number of cases of each area within the MLC with its expected number of cases such that the ratio between the number of cases and the population within the MLC is the same as that outside of it. More specifically, let c_i' be the updated number of cases of tract i within the MLC, then $c_i' = n_i(C - c_{\mathrm{MLC}})/(N - n_{\mathrm{MLC}})$, where C is the total number of cases, c_{MLC} is the number of cases within the MLC, n_i is the population of tract i, N is the total population and n_{MLC} is the population within the MLC. Conditional on the updated population and on the number of cases, the standard spatial scan statistic is applied to test the significance of the MLC within the updated data. This procedure is iterated until the MLC in the latest updated data is not significant.

Either of these two methods will be valid only if the type I error probability is close to the nominal level. In the next section, that is evaluated.

4. Type I Error and Power

A publicly available simulated benchmark dataset is used to evaluate the adjusted spatial scan statistic. Described in detail by Kulldorff et al. [10], it can be downloaded from "http://www.satscan.org/datasets".

The datasets use the geographical locations and the 1990 female population of 245 counties in the northeastern United States. The simulated datasets distributed 600 disease cases among the counties according to a multinomial distribution with probabilities proportional to the product of the population times the relative risk. Under the null hypothesis, the relative risk is equal to one. Under the alternative hypothesis, these relative risks are larger than 1 in the areas belonging to true clusters imposed on the map.

There are three types of high relative risk clusters, composed by rural, urban, or by mixed population areas. They are located in different parts of the map. The rural cluster is centered on Grand Isle County in northern Vermont on the Canadian border, which is the county with the smallest population, surrounded by other rural counties. The urban cluster is centered on Manhattan in New York City, surrounded by other urban counties. The mixed cluster is centered on Pittsburgh in western Pennsylvania, a large city surrounded by mostly rural counties. The clusters differ also by their size, being comprised by 1, 4 or 16 counties. There are 9 different types of datasets obtained by the crossing of different cluster sizes and locations.

Depending on the cluster size, the relative risks of the clusters' counties varied from 1.53 to 2.73 in the urban case, and from 2.10 to 2.85 in the mixed case. In the rural case, the relative risks were equal to 3.90 and 7.05 for 16 and 4 counties in the cluster, and equal to 192.89 in the cluster with a single county. For further details, see the paper by Kulldorff et al. [10].

There is one additional dataset classification, those with one true disease cluster and those with two true disease clusters. The two cluster datasets were built with an urban/rural, urban/mixed or mixed/rural combination, with the same number of counties in each cluster. While the number of counties is the same in both clusters, the relative risks and the population sizes are different. To calculate type I error probability of the adjusted spatial scan statistics for testing the significance of the second most likely cluster, benchmark data with one true disease cluster are needed. Each one of the type I error probabilities estimates is based on 10,000 simulated datasets. To calculate its power, benchmark data with two true disease clusters are needed. Each power estimate is based on 1000 simulated datasets.

Table 1 shows the estimated type I error probability of detecting a second cluster when there is in fact only one true cluster in the data. The tests were carried out at the nominal levels $\alpha = 0.01$ and $\alpha = 0.05$ with 0, 2, 5, 10 or 50 neighboring tracts removed, named as buffer size.

Generally, the estimated actual error type I probabilities are very close to the nominal levels. Table 1 shows that changing the buffer size has very little effect on the error type I probability. In the $\alpha = 0.01$ case, the maximum difference between actual and nominal levels is 0.001 for the mixed cluster and 0.003 for the urban cluster. In the $\alpha = 0.05$ case, these values are 0.006 and 0.004, respectively. Furthermore, the rural cluster with 4 and 16 counties do not have substantial differences between actual and nominal levels of significance. The only major discrepancies occur for the rural cluster composed by a single county, Grand Isle.

Table 1: Estimated type I error probabilities when testing the second most likely cluster when there is only one true cluster present, using the most likely cluster removing method. The buffer size is the number of neighboring counties removed together with the true cluster. The true clusters consist of 1, 4 or 16 counties located in an urban, rural or mixed population area, respectively, as described in the text.

Buffer size	Controlled	Urban			Rural			Mixed		
	α	1	4	16	1	4	16	1	4	16
0	0.01	0.007	0.007	0.009	0.017	0.010	0.009	0.010	0.011	0.010
2	0.01	0.007	0.007	0.009	0.016	0.008	0.009	0.011	0.009	0.009
5	0.01	0.008	0.008	0.009	0.016	0.010	0.009	0.009	0.010	0.009
10	0.01	0.011	0.007	0.008	0.016	0.009	0.010	0.010	0.009	0.009
50	0.01	0.010	0.010	0.009	0.008	0.010	0.010	0.010	0.009	0.011
0	0.05	0.047	0.043	0.042	0.059	0.052	0.048	0.047	0.049	0.051
2	0.05	0.044	0.044	0.042	0.059	0.051	0.049	0.050	0.047	0.050
5	0.05	0.045	0.045	0.040	0.058	0.052	0.048	0.048	0.047	0.050
10	0.05	0.053	0.048	0.040	0.057	0.051	0.050	0.050	0.048	0.052
50	0.05	0.054	0.047	0.045	0.053	0.048	0.051	0.051	0.052	0.056

Without a buffer, the estimated type I error probabilities are 0.017 and 0.059 for the nominal levels of 0.01 and 0.05 respectively, and the results are similar with a buffer added, except for the very largest buffer size.

To further investigate these type I error probabilities, we generated random dataset using different relative risks for the Grand Isle cluster, in addition to the original RR = 192.89. Also, in addition to the standard maximum window size of 50 percent of the population at risk size, we also estimated the type I error probabilities for a spatial scan statistic with a maximum window size of 5 percent. As shown in Table 2, the type I error probabilities are good when RR = 100 but slightly on the high side for the larger relative risks. The reason that the numbers for RR = 192.89 and circle size 50 percent is different from Table 1 is that a different random seed was used.

In Table 3, we show the estimated type I error probabilities using the method that replaces the observed counts within the MLC by the expected number of cases. The results are similar to those in Table 1, but not quite as good. In addition to the single Grand Isle cluster, there are also relatively large differences for all the rural and mixed clusters.

Table 4 shows the estimated power of the most likely cluster removing method, together with the standard scan statistic. The power of the latter is generally lower, with the biggest differences seen for larger cluster sizes. The power of the adjusted scan statistic power is not overly sensitive to the buffer size.

5. Breast Cancer Mortality in Northeastern United States

We illustrate the practical use of the adjusted spatial scan statistic using breast cancer mortality data for Northeastern United States, during 1988–1992, as collected by the National Center for Health Statistics. The study area composed by 245 counties and county equivalents in Northeast states including Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, Delaware, Maryland, and the District of Columbia. During 1988–1992, there were 58,943 breast cancer deaths recorded among women in this region, for which we have information about age and county of

Table 2: The estimated type I error probabilities for the second most likely cluster when there is only one true cluster in Grand Isle county, with 95% confidence intervals. Different relative risks were used, two different bounds on the maximum window size as well as two different buffer sizes defining the number of neighboring counties that were removed together with the counties in the most likely cluster detected. The circle size is the maximum geographic size of the scanning window defined in terms of a percentage of the total populations.

Relative risk	Buffer size	Circle size	$\alpha = 0.01$	$\alpha = 0.05$
100	0	5	0.010 (0.008,0.012)	0.043 (0.039,0.047)
100	0	50	0.013 (0.011,0.015)	0.045 (0.041,0.049)
100	50	5	0.010 (0.008,0.012)	0.047 (0.043,0.051)
100	50	50	0.008 (0.006,0.010)	0.051 (0.047,0.055)
192.89	0	5	0.014 (0.012,0.016)	0.056 (0.051,0.061)
192.89	0	50	0.014 (0.012,0.016)	0.057 (0.052,0.062)
192.89	50	5	0.013 (0.011,0.015)	0.054 (0.050,0.058)
192.89	50	50	0.010 (0.008,0.012)	0.058 (0.053,0.063)
200	0	5	0.012 (0.010,0.014)	0.057 (0.052,0.062)
200	0	50	0.013 (0.011,0.015)	0.057 (0.052,0.062)
200	50	5	0.013 (0.011,0.015)	0.052 (0.048,0.056)
200	50	50	0.010 (0.008,0.012)	0.056 (0.051,0.061)
400	0	5	0.011 (0.009,0.013)	0.055 (0.051,0.059)
400	0	50	0.010 (0.008,0.012)	0.051 (0.047,0.055)
400	50	5	0.010 (0.008,0.012)	0.052 (0.048,0.056)
400	50	50	0.010 (0.008,0.012)	0.056 (0.051,0.061)

Table 3: Estimated type I error probabilities when testing the second most likely cluster when there is only one true cluster present, using the most likely cluster repalcement method. The buffer size is the number of neighboring counties. The true cluster consisting of 1, 4 or 16 counties is located in an urban, rural or mixed population area respectively, as described in the text.

Buffer size	Controlled		Urban			Rural			Mixed	
	α	1	4	16	1	4	16	1	4	16
0	0.01	0.008	0.009	0.010	0.018	0.012	0.011	0.011	0.012	0.015
5	0.01	0.008	0.009	0.010	0.018	0.011	0.013	0.011	0.013	0.017
10	0.01	0.009	0.010	0.009	0.019	0.012	0.014	0.012	0.014	0.017
0	0.05	0.047	0.054	0.046	0.062	0.057	0.058	0.058	0.055	0.071
5	0.05	0.048	0.054	0.047	0.065	0.057	0.063	0.055	0.058	0.076
10	0.05	0.050	0.054	0.050	0.066	0.060	0.069	0.060	0.064	0.082

residence. The population is available for each county by five-year age groups. The total female population in the region was 29,535,210. The geographical location of each county was used as specified by the 1990 Census [11]. The data has been described in detail elsewhere [1]. The analysis is adjusted for age using indirect standardization.

Using the standard spatial scan statistic, it has previously been shown that women who live in the New York City - Philadelphia metropolitan area have an increased risk of dying from breast cancer compared to the rest of the region with 24044 cases when 23040 were expected [1]. This cluster is statistically significant ($P = .001$) and it is shown in Figure 1. Using the standard method, the second most likely cluster in the Buffalo area is not significant

Table 4: Comparison of power to test the second most likely cluster for significance levels 0.01 and 0.05 when two true clusters exist and the tests are the buffer removing method and the standard scan statistic method. For the standard method, power is based on the critical values under the the null hypothesis that there are no true clusters. The upper 0.01 and 0.05 log likelihood ratio critical values are 9.717 and 7.907 respectively.

Buffer size	Controlled	Rural Urban			Mixed Rural			MixedUrban		
	α	1	4	16	1	4	16	1	4	16
0	0.01	0.830	0.726	0.717	0.867	0.796	0.797	0.681	0.654	0.679
2	0.01	0.824	0.711	0.717	0.851	0.786	0.799	0.701	0.653	0.672
5	0.01	0.825	0.724	0.723	0.853	0.792	0.802	0.698	0.653	0.666
10	0.01	0.840	0.730	0.727	0.872	0.798	0.804	0.712	0.653	0.659
50	0.01	0.811	0.721	0.696	0.847	0.798	0.805	0.698	0.629	0.627
Standard	0.01	0.777	0.668	0.480	0.832	0.737	0.706	0.568	0.398	0.222
0	0.05	0.909	0.859	0.866	0.932	0.902	0.911	0.844	0.823	0.840
2	0.05	0.906	0.855	0.863	0.923	0.901	0.910	0.847	0.815	0.836
5	0.05	0.909	0.860	0.861	0.925	0.901	0.911	0.845	0.818	0.834
10	0.05	0.907	0.856	0.863	0.925	0.902	0.911	0.849	0.817	0.833
50	0.05	0.917	0.859	0.849	0.933	0.906	0.914	0.845	0.813	0.812
Standard	0.05	0.900	0.704	0.685	0.914	0.860	0.845	0.766	0.599	0.440

(P = .12), with 1416 observed deaths when 1280 were expected. This cluster is shown in Figure 1, in the northwestern part of the map. Using the sequential method, with the most likely cluster removal method without a buffer, the second most likely cluster is instead in the Boston metropolitan area (P = .0001), with 5966 observed deaths when 55565 were expected.

A complete list of the detected clusters is presented in Table 5, for both the standard and the adjusted spatial scan statistics. When using the standard method, only the most likely cluster was statistically significant. When adjusting for more likely clusters using the sequential method, six statistically significant clusters were found at the $\alpha = 0.05$ level. This illustrates the behavior of the usual scan statistic which, by not taking into account the previously detected clusters, may miss some less likely clusters, and especially if the most likely cluster is large in size. Compare, for example, the P-values for the cluster around Boston, ranked as fourth in the standard scan statistic with $P = .40$ and as second in the sequential method with $P = .0001$. Note also that two significant clusters found by the sequential method, Pittsburgh and Albany, are not among the five most likely clusters found by the standard spatial scan statistic.

6. Discussion

When using the standard spatial scan statistic, a large cluster in one area of the map may hide the existence of a secondary cluster in another area of the map. By removing the MLC, secondary clusters can be detected using the standard scan statistic on the reduced dataset. We have shown that this procedure has an actual error type probability very close to the nominal level and that it has a higher power than the standard scan statistic for secondary clusters. The sequential approach can be applied recursively to locate other clusters conditionally on the presence of all the previously detected clusters. This is repeated until the most likely cluster in the updated data is nonsignificant. For the breast cancer

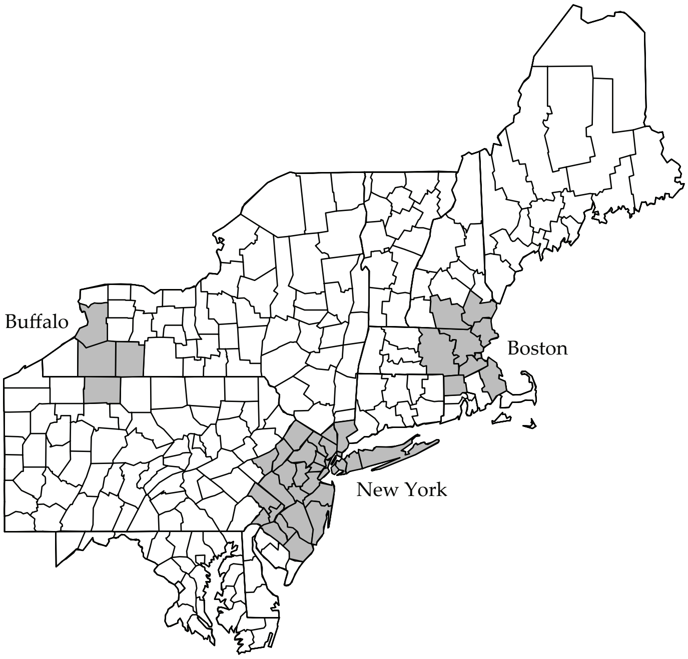

Figure 1: The most likely cluster and the second most likely cluster using the standard scan method and the sequential scan method. The most likely cluster is around New York City, the second most likely cluster detected using the standard method is close to Buffalo and it is not significant. The second most likely cluster detected using the sequential method is around Boston and it is significant.

Table 5: Breast cancer mortality analysis for women in the Northeast United States, 1988–1992, using the standard and the sequential spatial scan statistics. RR = relative risk within the cluster compared to the rest of the Northeast; LLR: log likelihood ratio.

Method	Cluster	Location	Counties	Deaths	Expected	RR	LLR	*P*-value
Standard	1	NYC-Philadelphia	32	24044	23040	1.07	35.7	0.0001
	2	Buffalo	4	1416	1280	1.11	7.1	0.12
	3	Washington DC	1	712	618	1.15	6.9	0.15
	4	Boston	9	5966	5726	1.05	5.5	0.40
	5	Eastern Maine	3	267	229	1.17	3.0	0.99
Sequential	1	NYC-Philadelphia	32	24044	23040	1.07	35.7	0.0001
	2	Boston	9	5966	5565	1.07	16.8	0.0001
	3	Buffalo-Rochester	9	2362	2119	1.12	14.5	0.0004
	4	Baltimore-Washington	17	4255	3901	1.09	18.3	0.0001
	5	Pittsburgh	1	1765	1565	1.13	13.3	0.0002
	6	Albany	19	2336	2156	1.08	8.2	0.04
	7	Eastern Maine	3	267	210	1.27	7.2	0.07

mortality data from the Northeastern United States, we make different inference about the presence of disease clusters when using the two methods. While we find only one significant most likely cluster using the standard method we find six significant clusters using the sequential approach. This may have important practical implications.

A key finding of this study is that when adjusting the analysis for more likely clusters, it is better to remove these clusters rather than replacing them with expected counts. While the type I error probabilities are close to the nominal levels for both approaches, they are consistently closer when the removing method is used. Another important finding is that there is no need to use a buffer around the clusters to be removed, but it is enough to remove the cluster itself. The type I error probabilities are about equally good in both cases, while the power suffers slightly when a large buffer is used.

If there is a reliable external estimate for the expected counts under the null hypothesis, and we are interested in the absolute differences in risk compared to this external estimate rather than the geographical variation of the relative risks within the study region, we should use the unconditional rather than the conditional scan statistic. Sonesson [8] has shown that this will give higher statistical power. For the unconditional scan statistics there is no need to use a sequential approach though, as the expected number of cases will not change when the most likely cluster is removed. The only effect of the sequential approach would be a slight change in the P-value of secondary clusters due to slightly less multiple testing that needs to be adjusted for. Hence, we have only considered the conditional scan statistic in this paper.

While we used a sequential approach for the breast cancer mortality example, by removing more likely clusters when evaluating the statistical significance of clusters with lower likelihood ratios, the procedure could also be used to evaluate the stronger clusters after adjusting for weaker ones. For example, one could check whether the Albany cluster has a much lower P-value than its current .04 after removing and therefore adjusting for the Eastern Maine cluster. In this manner, one may evaluate the statistical significance of any cluster adjusting for any collection of other clusters. The P-values will of course be different for the same cluster depending on what other clusters were adjusted for, just like in a regression analysis, where the P-value of one variable will depend on what other variables are included (adjusted for) in the regression model.

The spatial scan statistic is commonly used to analyze mortality or incidence data using a Poisson model, but there is also for example a Bernoulli model for dichotomous case-control data and an exponential model for survival data with or without censoring [12]. If one is not willing to assume a parametric distribution to the observed counts, an alternative is the semiparametric method of Wen and Kedem [13] which provides inference for the primary and secondary clusters by means of a false discovery rate method.

The sequential scan statistic approach may also be applied for these other probability models to adjust for multiple clusters. The same may or may not be true for other types of scan statistics such as purely temporal and space-time scan statistics, whether performed in a retrospective or prospective manner [14, 15], or scan statistics that are designed to detect areas with lower rather than higher rates of disease, and we do not know how well a sequential procedure would work in such a setting. Moreover, while the scan statistic is most often applied using a circular window as we did in this paper, there have more recently been developments of nonparametric shaped scan statistics as well [16–19], in addition to other noncircular shapes [20]. It would be interesting to evaluate how the sequential approach works in these settings, to learn whether the type I error probabilities are still under control and close to their nominal values.

The sequential scan statistic presented here is more computer intensive than the standard scan statistic since a standard analysis must be run for each sequential stage of the procedure. The sequential scan statistic has been incorporated as an option into the freely available SaTScan software (http://www.satscan.org/) that already contain the standard temporal, spatial and space-time scan statistics.

Acknowledgment

This research was funded by the National Cancer Institute, Grant no. RO1CA095979, and by the National Institute for Child Health and Development, Grant no. R01HD048852.

References

[1] M. Kulldorff, E. J. Feuer, B. A. Miller, and L. S. Freedman, "Breast cancer clusters in the northeast United States: a geographic analysis," *American Journal of Epidemiology*, vol. 146, no. 2, pp. 161–170, 1997.

[2] C. E. Hsu, H. Jacobson, and F. S. Mas, "Evaluating the disparity of female breast cancer mortality among racial groups—a spatiotemporal analysis," *International Journal of Health Geographics*, vol. 3, article 4, 2004.

[3] A. Odoi, S. W. Martin, P. Michel, D. Middleton, J. Holt, and J. Wilson, "Investigation ofclusters of giardiasis using GIS and a spatial scan statistic," *International Journal of Health Geographics*, vol. 3, article 11, 2004.

[4] A. L. Andrade, S. A. Silva, C. M. Martelli et al., "Population-based surveillance of pediatric pneumonia: use of spatial analysis in an urban area of Central Brazil," *Cadernos de Saúde Pública*, vol. 20, no. 2, pp. 411–421, 2004.

[5] C. H. Washington, J. Radday, T. G. Streit et al., "Spatial clustering of filarial transmission before and after a Mass Drug Administration in a setting of low infection prevalence," *Filaria Journal*, vol. 3, p. 3, 2004.

[6] R. Heffernan, F. Mostashari, D. Das, A. Karpati, M. Kulidorff, and D. Weiss, "Syndromic surveillance in public health practice, New York City," *Emerging Infectious Diseases*, vol. 10, no. 5, pp. 858–864, 2004.

[7] J. Glaz, J. Naus, and S. Wallenstein, *Scan Statistics*, Springer Series in Statistics, Springer, New York, NY, USA, 2001.

[8] C. Sonesson, "A CUSUM framework for detection of space-time disease clusters using scan statistics," *Statistics in Medicine*, vol. 26, no. 26, pp. 4770–4789, 2007.

[9] M. Dwass, "Modified randomization tests for nonparametric hypotheses," *Annals of Mathematical Statistics*, vol. 28, pp. 181–187, 1957.

[10] M. Kulldorff, T. Tango, and P. J. Park, "Power comparisons for disease clustering tests," *Computational Statistics & Data Analysis*, vol. 42, no. 4, pp. 665–684, 2003.

[11] US Bureau of the Census, *Statistical Abstracts of the United States*, GPO, Washington, DC, USA, 111th edition, 1991.

[12] L. Huang, M. Kulldorff, and D. Gregorio, "A spatial scan statistic for survival data," *Biometrics*, vol. 63, no. 1, pp. 109–312, 2007.

[13] S. Wen and B. Kedem, "A semiparametric cluster detection method—a comprehensive power comparison with Kulldorff's method," *International Journal of Health Geographics*, vol. 8, pp. 73–88, 2009.

[14] M. Kulldorff, "Prospective time periodic geographical disease surveillance using a scan statistic," *Journal of the Royal Statistical Society. Series A*, vol. 164, no. 1, pp. 61–72, 2001.

[15] M. Kulldorff, R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari, "A space-time permutation scan statistic for disease outbreak detection," *PLoS Medicine*, vol. 2, pp. 216–224, 2005.

[16] G. P. Patil and C. Taillie, "Upper level set scan statistic for detecting arbitrarily shaped hotspots," *Environmental and Ecological Statistics*, vol. 11, no. 2, pp. 183–197, 2004.

[17] L. Duczmal and R. Assuncáo, "A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters," *Computational Statistics and Data Analysis*, vol. 45, no. 2, pp. 269–286, 2004.

[18] R. Assunção, M. Costa, A. Tavares, and S. Ferreira, "Fast detection of arbitrarily shaped disease clusters," *Statistics in Medicine*, vol. 25, no. 5, pp. 723–742, 2006.

[19] T. Tango and K. Takahashi, "A flexibly shaped spatial scan statistic for detecting clusters," *International Journal of Health Geographics*, vol. 4, pp. 4–11, 2005.

[20] L. E. Christiansen, J. S. Andersen, H. C. Wegener, and H. Madsen, "Spatial scan statistics using elliptic windows," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 11, no. 4, pp. 411–424, 2006.