

Spatial Scan Statistics Using Elliptic Windows

L. E. CHRISTIANSEN, J. S. ANDERSEN, H. C. WEGENER, and H. MADSEN

The spatial scan statistic is widely used to search for clusters. This article shows that the usually applied elimination of secondary clusters as implemented in SatScan is sensitive to smooth changes in the shape of the clusters. We present an algorithm for generation of a set of confocal elliptic windows and propose a new way to present the information when a spatial point process is considered. This method gives smooth changes for smooth expansions of the set of clusters. A simulation study is used to show how the elliptic windows outperforms the usual circular windows. The proposed method for graphical representation of the information in a set of clusters contain more information than just presenting nonoverlapping clusters. We suggest that more than one graphical representation of a set of clusters should be used to easily extract more information and to avoid pitfalls of the selected method.

Key Words: Broilers; *Campylobacter*; Likelihood ratio; Monte Carlo; Overlapping clusters; SatScan.

1. INTRODUCTION

In epidemiology the spatial scan statistic as implemented in SaTScan (www.SaTScan.org) is widely used to search for clusters in spatial data. The software has been used for many applications including infectious diseases (Sauders et al. 2003), cancer (Roche, Skinner, and Weinstein 2002), and veterinary epidemiology (Enemark et al. 2002; Hoar et al. 2003).

SaTScan is used in cases with few spatial locations as well as in cases with many locations.

In the case where secondary clusters are of interest SaTScan eliminates all clusters overlapping with the most significant cluster; this gives a large reduction in the number of significant clusters and loss of information—especially if a case with hundreds of spatial locations is considered.

Not all naturally occurring phenomena are circular and hence there is a need for methods that can find clusters with other shapes.

L. E. Christiansen is Post Doc., Informatics and Mathematical Modelling, The Technical University of Denmark, Denmark (E-mail: lec@imm.dtu.dk). J. S. Andersen is Senior Statistician, Group Clinical Development, ALK-Abello A/S, Denmark. H. C. Wegener is Head of Institute, Institute of Food Safety, Institute for Danish Food and Veterinary Research, Denmark. H. Madsen is Professor, Informatics and Mathematical Modelling, The Technical University of Denmark, Denmark.

© 2006 American Statistical Association and the International Biometric Society
Journal of Agricultural, Biological, and Environmental Statistics, Volume 11, Number 4, Pages 411–424
DOI: 10.1198/108571106X154858

This article presents an algorithm for construction of a set of confocal ellipses. A simulation study shows that even when the true underlying cluster is circular the use of elliptic windows performs better than using circular windows. Finally, an example illustrates a new way to present the information in a set of clusters found using either circular or elliptic scanning windows.

2. MODEL DESCRIPTION

The spatial scan statistic is a likelihood ratio statistic for spatial point processes. Originally the spatial scan statistic was a 1D statistic, but it has been developed to manage both 2D and 3D problems (Kulldorff 1997, 1999). The 1D version was designed for temporal data, the higher dimensional designs were made for spatial and for combinations of spatial and temporal data. The implementation in SaTScan includes methods for both Bernoulli and Poisson point processes with circular windows of varying size. Improvements of the algorithm have been suggested by several researchers; for example, Boscoe, McLaughlin, Schymura, and Kielb (2003) presented a scheme that does the elimination on 10 levels of significance and Gangnon and Clayton (2001) worked with a weighted average likelihood ratio test.

The present work concerns only the Bernoulli model but the comments apply to the Poisson model as well. Let a geographical region G be the area of interest. Within this region there are a number of points each having a population of one or more cases or non-cases. The total population summed over all points is n_G and includes c_G cases. Any subset of points in G is a potential cluster. Let $Z \subset G$ contain a population of n_Z observations including c_Z cases.

Let \mathcal{Z} be the set of subsets of a given family, for example, set of confocal ellipses (a description of the generation is postponed until Section 3).

At first high-risk clusters are considered. The null hypothesis is that the incidence rate is the same all over the study region and the alternative hypothesis is that the rate is higher in Z . Under the null hypothesis the maximum likelihood is given by

$$L_0(G) \propto \left(\frac{c_G}{n_G}\right)^{c_G} \left(1 - \frac{c_G}{n_G}\right)^{n_G - c_G}.$$

Under the alternative hypothesis the maximum likelihood for the subset Z being a high-risk cluster is:

$$L(Z) \propto \left(\frac{c_Z}{n_Z}\right)^{c_Z} \left(1 - \frac{c_Z}{n_Z}\right)^{n_Z - c_Z} \left(\frac{c_G - c_Z}{n_G - n_Z}\right)^{c_G - c_Z} \left(1 - \frac{c_G - c_Z}{n_G - n_Z}\right)^{n_G - c_G - (n_Z - c_Z)}$$

if $c_Z/n_Z > (c_G - c_Z)/(n_G - n_Z)$ and

$$L(Z) \propto \left(\frac{c_G}{n_G}\right)^{c_G} \left(1 - \frac{c_G}{n_G}\right)^{n_G - c_G}$$

otherwise. If low-risk clusters are of interest one uses non-cases instead of cases as the observations. If the interest is in both high- and low-risk clusters simultaneously, $c_Z/n_Z >$

$(c_G - c_Z)/(n_G - n_Z)$ determines if a subset is to be considered a high-risk subset (true) or a low-risk subset (false).

Next, the subset with the highest maximum likelihood is

$$\hat{Z} = \{Z : L(Z) \geq L(Z') \forall Z' \in \mathcal{Z}\}.$$

This is the most likely cluster and statistical inference is done using the likelihood ratio (LR):

$$\text{LR} = \frac{L(\hat{Z})}{L_0(G)}.$$

The distribution of the likelihood ratio is not available in a closed analytical form. Instead Monte Carlo simulation is used to generate a distribution of the likelihood ratios under the null hypothesis. See Section 4 for further details on the Monte Carlo simulations.

3. GENERATION OF SUBSETS

In SaTScan concentric circles are used to construct a set of subsets. This article presents a new algorithm for generation of confocal ellipses leading to a larger set of subsets and a set that includes the set of concentric circles as a true subset. The motivation for this larger and more flexible set is that not all naturally occurring phenomena can be assumed to be circular.

First restating the algorithm used in SaTScan to generate circular subsets. In principle it is possible to make an infinite set of circles in the study region, but in practice only different subsets of the locations in the point process are of interest. Thus for a given center an increase in radius is only of interest if a new point enters the circle. Having n locations and setting the maximum cluster size to 50% leads to a maximum of $0.5 \cdot n$ subsets per center. Centers that are too close give rise to the same subsets and are not of interest. In SaTScan one can either specify a list of centers or use the locations of the points in the point process as centers. Here the latter was chosen.

An algorithm for generation of concentric circles including up to a prespecified proportion of the locations, say 0.50 is as follows:

1. *Select one point that will be the center and calculate the distance from this point to all other points.*
2. *Sort the points according to the distance to the selected center.*
3. *Make the point at the selected center the first subset.*
4. *Include the nearest point and make it the next subset.*
5. *Keep including the next nearest point, until the prespecified proportion is included, making a new subset for every inclusion.*
6. *Repeat 1 through 5 making all points a center one by one.*

7. *Remove identical subsets.*

A larger and more flexible but still finite family of subsets consist of concentric ellipses. The foci are chosen among the locations in the point process. This makes the maximum length of the constructed set of subsets $n^3 p_m p_s$, where p_m is the proportion of the total number of locations in the largest cluster, and p_s is the proportion of the total number of locations that are used as the secondary foci.

An algorithm for generation of confocal ellipses including up to a prespecified proportion of the locations, say 0.50 is as follows:

1. *Select one point, which will be the first focal and calculate the distance from this point to all other points.*
2. *Sort the points according to the distance to the first focal.*
3. *Choose the first focal as the second focal in order to start with the set of circles. Include the next nearest point, until the prespecified proportion is included, making a new subset for every inclusion.*
4. *Make the next nearest point the second focal and calculate the distance from this point to all other points.*
5. *Sort the points according to the sum of the distances to the two foci.*
6. *Include the nearest point and make it the next subset.*
7. *Keep including the next nearest point, until the prespecified proportion is included, making a new subset for every inclusion.*
8. *Repeat 4 through 7 until the prespecified proportion of secondary foci has been used.*
9. *Repeat 1 through 8 making all points the first focus, one at a time.*
10. *Remove identical subsets.*

The removal of identical subsets can be optimized by adding Step

- 4b. *If the two foci were previously used as foci go to Step 4.*

This reduces the number of generated subsets in Step 1 through 9 by up to 50% dependent on the mutual distances, that is, dependent on changes in the density of locations.

The choice of p_m and p_s depends on the nature of the phenomena under analysis. Choosing $p_m = 0.50$ seems reasonable in many cases as a cluster should not take up more than half the space. From the simulation study in Section 5.1 $p_s = 0.2$ seems to be reasonable.

4. MAKING INFERENCE USING MONTE CARLO

The previous section describes two different algorithms for the generation of a set of subsets. Furthermore, an expression is derived for the likelihood ratio associated with each subset. The most likely cluster among the subsets is the one with the highest likelihood ratio.

In the general setting it is infeasible, if at all possible, to write an expression for the distribution of likelihood ratios in closed analytical form. Hence, in order to test if the null hypothesis holds Monte Carlo simulations are used. The Monte Carlo method outlined in the following was presented by Kulldorff (1997). One Monte Carlo simulation is made by generating a new dataset under the null hypothesis, that is, a point process with the same number of observations at the same locations and with the same risk of being a case as the over all risk in the original dataset. The most likely cluster is found using the simulated data and recording its likelihood ratio. To make inference on a 1% significance level using 9,999 simulations the most likely subset is significant if its likelihood ratio is among the 99 highest Monte Carlo maximum likelihood ratios (MCMLR).

If the null hypothesis is rejected for the most likely subset, then all the other subsets having one additional point or lacking one point when compared with the most likely subset will in most cases have likelihood ratios just below the most likely one; this is to recall that the underlying cluster may be of a different size. It is also of interest if there are other significant subsets representing totally different clusters. When testing other clusters than the most likely do keep in mind that the Monte Carlo generated distribution is made of the most likely cluster in the simulated datasets. The second most likely cluster should be tested against the second most likely in the simulated datasets, and so forth. This is not feasible, but the fact that the second most likely cluster has a lower likelihood ratio than the most likely cluster makes it possible to make an approximate test. When using the distribution for the most likely cluster to test secondary clusters the true significance level is greater than or equal to the reported significance level.

In a case with many subsets having a likelihood ratio above the significance level it becomes difficult to present the information. There are several possible solutions to this problem, the one implemented in SaTScan starts by eliminating all subsets having a nonempty intersection with the most likely subset. Then all subsets having a nonempty intersection with the second most likely subset are eliminated and so forth. This makes a dramatic reduction in the number of subsets. We propose an alternative: For each location the proportion of significant subsets that it belongs to is used as a measure (could be weighted by the likelihood ratios). In the case where both high- and low-risk areas are of interest these are treated separately and care should be taken to make it possible to see if some locations are both included in high- and low-risk subsets. This gives a more detailed picture of the underlying cluster(s).

5. EXAMPLES

Two examples are presented, first a simulation study comparing the ability to detect underlying circular clusters using circular and elliptic scanning windows. Secondly, a dataset with *Campylobacter* infections in Danish broilers is used to illustrate the differences in acquired information using two different schemes to present significant clusters as well as the differences between circular and elliptic clusters.

5.1 SIMULATION STUDY

A simulation study was used to compare the capability to find the correct circular cluster using either circular or elliptic scanning windows. For each simulation a dataset consisting of 50 points randomly distributed in a 2D unit box was made, that is, $x, y \in U([0; 1])$. All points within 0.25 from the center of the box are within the cluster, that is, a centered circle covering 20% of the area. At all points the population is uniformly distributed between 5 and 20 and the number of cases is found using a binomial distribution with probability parameter p_{in} inside the cluster and p_{out} outside the cluster. For each such dataset the most likely circular and elliptic cluster was found and the number of points from the true cluster was recorded together with Type 1 and Type 2 errors for the circular and elliptic clusters. Here, a Type 1 error is rejecting a true hypothesis, that is, the number of points inside the true cluster that was not detected; a Type 2 error is accepting a false hypothesis, that is, the number of points that were detected but does not belong to the true cluster.

As each simulation does not have the same number of points in the true underlying cluster the proportions of Type 1 and Type 2 errors were reported. Figure 1 shows the results of two sets of experiments, the first investigates the dependence on p_s and the second the effect of changing the difference between p_{in} and p_{out} . For each setting 10,000 datasets are simulated and the full simulation and estimation time for all the 400,000 simulations was 200 minutes on a 1.2GHz SUN UltraSPARC3 CPU (approximately the same on a 1.1GHz Intel P3 laptop).

The result of the first experiment, changing p_s can be seen in Figure 1(a) and (b) showing Type 1 and Type 2 errors, respectively. Both for circular scanning windows (black lines) and elliptic scanning windows (blue lines) the mean, 75% and 90% fractiles are found for each value of p_s . As the circular windows does not depend on p_s the black lines should be horizontal and so they are when neglecting the small disturbances mostly seen for the Type 1 error—these are due to the discrete space of ratios between small integers and doing more repetitions will remove this. It is seen that there is only difference for $p_s < 0.1$ and that the error rate drops fast until $p_s = 0.2$ and then drops slowly. The mean Type 1 and 2 error is halved using $p_s \geq 0.3$ and $p_s \geq 0.25$, respectively.

In the second experiment p_s was fixed at 0.2 as larger values gives only small improvements at the cost of a large increase in the number of scanning windows. p_{out} was fixed at 0.4 and p_{in} was changed to find the sensitivity of the two windows. The results can be seen in Figure 1(c) and (d) where the same coding is used as in the first example. It is seen that when the $p_{\text{in}} < p_{\text{out}} + 0.1$ at least 25% of the most likely clusters does not include any

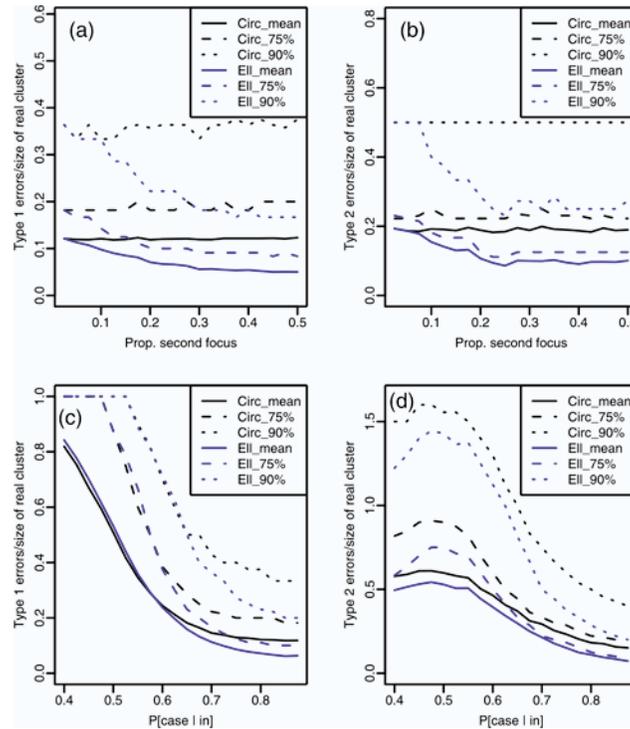


Figure 1. The result of a simulation study used to compare circular and elliptic scanning windows when the true cluster is circular. The probability of being a case is 0.40 outside the cluster, the probability of being a case inside the cluster was 0.80. Plot (a) and (b) show the effect of changing p_s on the proportion of Type 1 and Type 2 errors, respectively. Plot (c) and (d) show the effect of changing the probability of being a case, when inside the circular cluster, on the proportion of Type 1 and Type 2 errors, respectively. $p_s = 0.20$ and $p_{out} = 0.40$ were used. In all plots the mean and 75% and 90% quantiles are plotted for both circular and elliptic scanning windows.

of the points in the true cluster. When $p_{in} < 0.6$ the two scanning windows have almost the same Type 1 error rate, for larger p_{in} the elliptic scanning windows outperforms the circular scanning windows. When looking at the Type 2 error rate the elliptic windows are better for all the used values of p_{in} . One partial explanation is that the average size of the most likely elliptic clusters are smaller than for circular clusters.

One reason why elliptic windows are superior to circular windows when the true cluster is circular is that in most cases there will be no point in the center of the true cluster but often two points will be on opposite sites of the center and not too far from the center. Using such two points as foci for an ellipse and including additional points a better representation of the true circular cluster can be obtained than when using either of the foci as center for a circular window.

5.2 DATASET: *Campylobacter* IN DANISH BROILERS

The dataset is from a national monitoring program for Danish broiler flocks examined for *Campylobacter* by cloacal swabs at slaughter. All poultry flocks slaughtered in Denmark between 1998 and 2001 were in the program. During this period 23,279 broiler flocks were

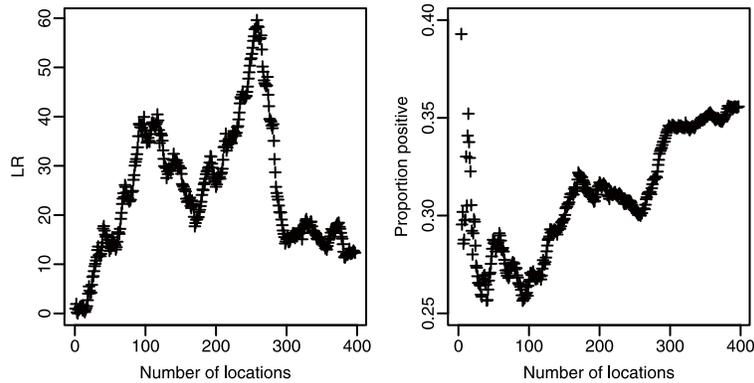


Figure 2. The likelihood ratio (left) and proportion of positive flocks (right) as a function of the number of locations in the concentric circles that has the same center as the most likely circle.

sampled; 7,741 flocks were excluded from the dataset either because they were reported to come from a farm not registered for broiler production or were reported to come from an unknown house. Another 7,217 flocks were excluded because they were not from the first batch. The reason for only including samples from the first batch is that the risk increases for the following batches and to avoid dependence between batches. This leaves 8,321 flocks in the dataset.

Only flocks originating in Jutland or at Funen are included; this reduces the number of houses from 828 to 794 and the number of flocks from 8,321 to 8,056. Houses on the same farm are considered lying on a row with 10 meters between houses. In total 3,080 of the flocks were positive corresponding to a 38.2% risk that a flock is infected.

The sampling was performed at the slaughterhouse at the entrance to the slaughterline by collecting one cloacal swab from 10 animals. If at least one of these ten swabs tested positive for *Campylobacter* spp. the flock was labeled positive.

The dataset was collected under a national monitoring program and has also been used to find climatic predictors for the pronounced seasonality observed in *Campylobacter* infections in broiler flocks (Patrick et al. 2004).

5.3 PRESENTING SIGNIFICANT CLUSTERS

The geographical locations in the dataset described above were used to generate circular clusters including up to 50% of the locations. Before removing duplets the full set of clusters contained 315,218 clusters. That set was reduced to 107,953 unique clusters. The next step was calculating the maximum likelihood ratios based on the dataset. It was chosen to use the 10,000 most likely clusters in our analysis, that is, the 10,000 highest maximum likelihood ratios.

In this case the most likely cluster is low prevalent, but why this cluster and not the one including one additional location or one less? The left part of Figure 2 shows the likelihood ratio for the circles that have the same center as the most likely cluster, it is seen that the clusters containing just about the same number of locations have about the same likelihood

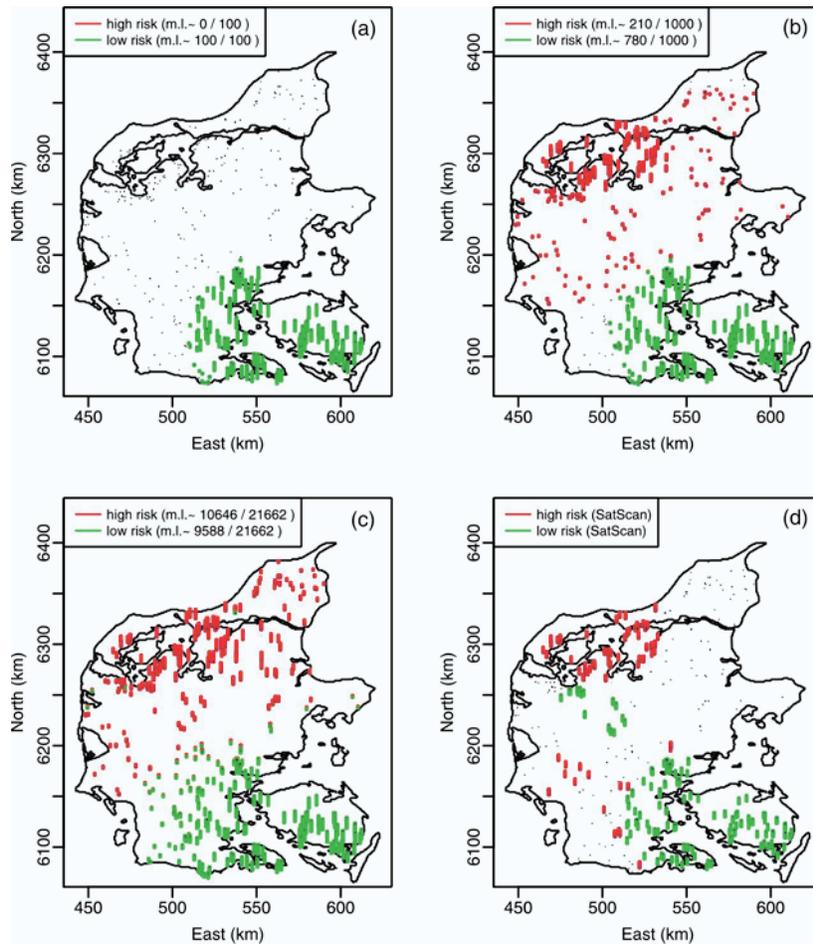


Figure 3. Plots (a) through (c) showing the geographical distribution of the 100, 1,000, and 21,662 most significant circular clusters, the 21,662 clusters are the ones with a significance of 0.0001 or less. Green is low prevalent and red is high prevalent, the length of the markers indicates the proportion of clusters that a given house is in (see the legends). The lower right plot is made using the output from SaTScan.

ratio. Therefore one has to be aware that the underlying cluster may be different from the most likely one. It is also noted that a range of clusters less than half the size have relatively high likelihood ratios. When looking in the right part one can see that the proportion of positive flocks is much lower in the region with about 100 locations when compared to the proportion of the most likely cluster. This indicates that there is a core of the most likely cluster that is accountable and hence those locations may be of greater interest than the most likely cluster. It is also seen that the likelihood ratio drops fast when the circles becomes larger than the most likely cluster.

Taking a look at the clusters: Figure 3 shows four maps of the region of interest with a dot at every broiler house and red and green bars showing the occurrence of locations in high- and low-risk clusters, respectively. The lower right plot shows the result of a run with SaTScan showing the clusters that are significant at a 5% level. The upper left plot is made

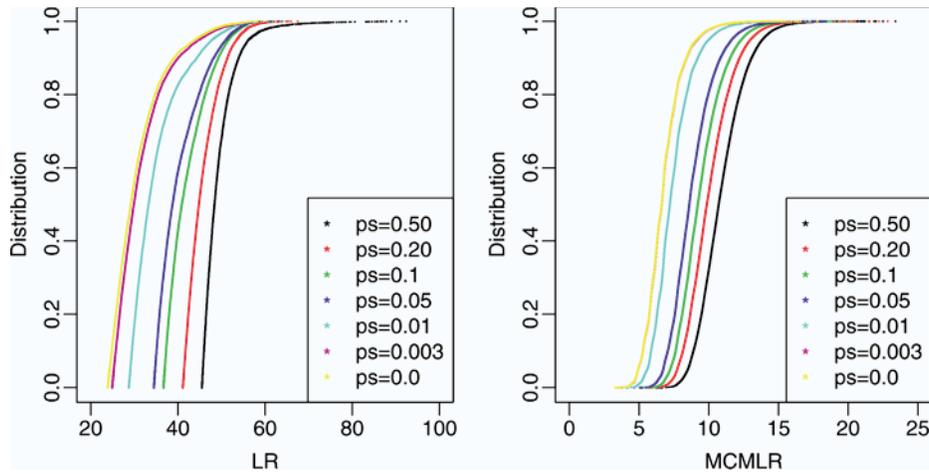


Figure 4. Distributions of the 10,000 highest likelihood ratios (left) and the 99,999 MCMLRs (right).

using the 100 most significant clusters without eliminating any clusters besides the duplets. The length of the bars indicate how large a proportion of the high- and low-risk clusters a location belongs to. The highest occurrence of high and low is indicated in the legend. This is the same for the second and third plots, except that they are made using 1,000 and 21,662 clusters, respectively. The 21,662 clusters are the circular clusters that are significant at a 10^{-4} level.

It is seen that the 100 most likely clusters all define the same low-risk area and this was also expected. When including 1,000 clusters some high-risk clusters are included, and these define an area in the northwestern part of Jutland and a few cover the middle part of Jutland. So far the areas covered by clusters are in agreement with the two most likely clusters found using SaTScan. This agreement fails when the 21,662 most significant clusters are used, as can be seen in the lower left subfigure. The low-risk area is almost stable but the high-risk clusters cover more and more. Some locations in the northern part of the low-risk area are included both in low- and high-risk clusters. When comparing with SaTScan one notes that SaTScan finds a low-risk area just south of the first high-risk cluster and some small high-risk clusters surrounding the most likely cluster. Those small high-risk areas appear as most other clusters are removed due to an overlap with the most likely cluster. So the selection done by SaTScan is in agreement with our representation as long as the clusters are far apart.

5.4 ELLIPTIC CLUSTERS

The next step is a comparison of circular and elliptic clusters. First take a look at Figure 4. The yellow distribution is made using circles including up to 50% of the locations. The other distributions are made using ellipses also limited to 50% of the locations and including up to the 0.3%, 1%, 20%, and 50% nearest locations as the second focus, that is, $p_m = 0.5$ in all cases and $p_s \in [0, 0.003, 0.01, 0.20, 0.50]$ in the specified order. In both parts of the

Table 1. Summary of Results for Eight Runs With Fixed p_m and Varying p_s . For all runs 99,999 Monte Carlo simulations are made, and thus a cluster is significant at a 10^{-5} level if its likelihood ratio is above the highest MCMLR.

p_m	p_s	No. clusters with duplets	No. clusters without duplets	Highest LR	Highest MCMLR
0.50	0.0	315,218	107,953	59.628	17.7875
0.50	0.003	537,374	126,708	59.628	16.689
0.50	0.01	1,525,790	223,623	59.628	19.9197
0.50	0.02	3,080,090	351,406	59.628	19.3834
0.50	0.05	7,662,602	717,338	63.1764	19.8396
0.50	0.10	15,134,726	1,308,741	64.0019	20.1673
0.50	0.20	29,791,478	2,655,066	67.4917	22.8339
0.50	0.50	70,288,418	7,939,616	92.3859	23.3868

figure a smooth transition is seen—more flexible shapes result in higher likelihood ratios. This is as expected.

The removal of duplets makes a large reduction in the number of clusters where the likelihood ratio has to be evaluated. Table 1 shows that the reduction is between a factor of 3 and 10, and that a larger reduction is seen for more flexible clusters, that is, ellipses with a high p_s . The predicted maximum length for the largest case with $p_m = p_s = 0.5$ is above 125 mio. clusters. This number reduces to less than 8 mio. clusters due to the removal of clusters thus speeding up the calculation by a factor of 15. $p_s = 0.003$ was chosen as this only allows the nearest neighbor as a secondary focus and this only results in a small increase in the number of unique clusters. In fact the most likely cluster is the same when $p_s < 0.025$. From the table it is also seen that the highest likelihood ratio is about three to four times higher than the highest MCMLR indicating that the clusters are highly significant in the present dataset. To emphasize the significance another test was performed with 9,999,999 Monte Carlo simulations, and in this case the highest MCMLR was 23.4981, so these clusters are extremely significant with the given test.

To investigate how the elimination implemented in SaTScan works, when applied to a more flexible set of clusters, all the clusters having a significance level above 10^{-4} were selected. The two plots in the top of Figure 5 show the clusters found with $p_s = 0.01$ (left) and 0.2 (right). These should be compared with Figure 3(d). It is seen that the change from circles to ellipses with $p_s = 0.01$ primarily changes the shape of the secondary clusters a little. When compared with the set of ellipses with $p_s = 0.2$ the picture is changed quite dramatically. The low-risk cluster on Funen and surroundings is almost stable but the southernmost locations in Jutland that used to be part of the cluster have now joined a high-risk cluster. The high-risk area in the northwestern part of Jutland is now split into two small high-risk clusters leaving the impression that it may be a localized phenomenon. The two bottom plots show the geographical distribution of all the clusters that are significant at a 10^{-4} level with $p_s = 0.01$ (left) and 0.2 (right). When compared with Figure 3(c) only small changes are seen, so the overall changes in the most likely clusters are small. In

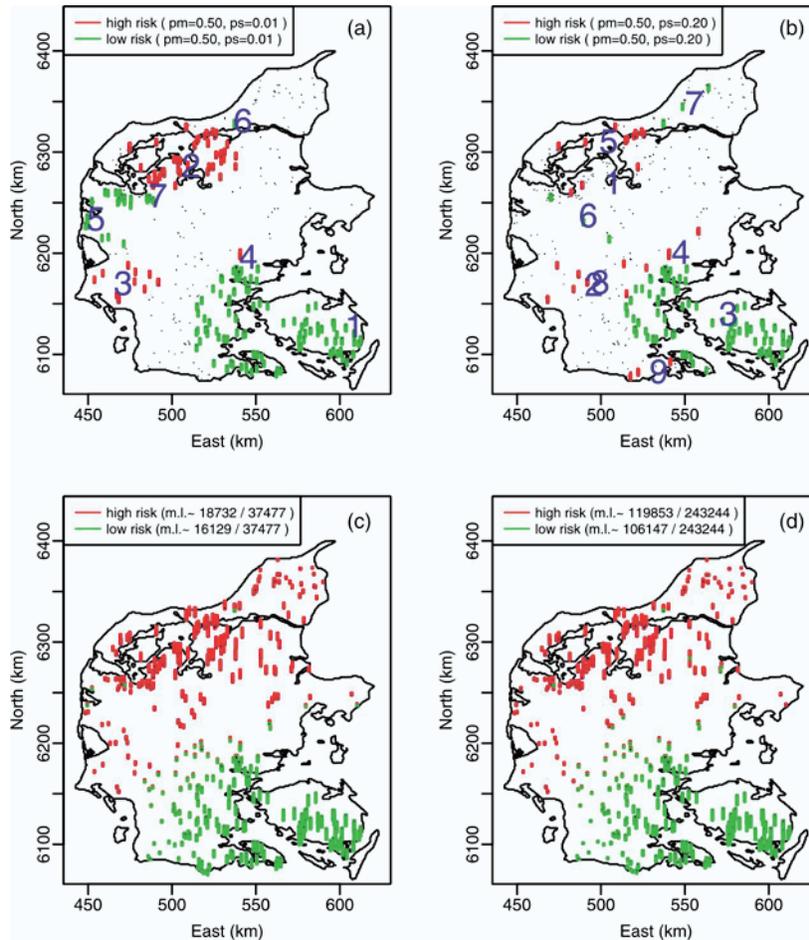


Figure 5. Plots presenting elliptic clusters that are significant at a 10^{-4} level. The two to the left are made with $p_s = 0.01$ and the two to the right with $p_s = 0.2$. The two at top show the result of SaTScan elimination and the two at the bottom show all the significant clusters.

the case with $p_s = 0.2$ a total of 243,244 clusters were significant at a 10^{-4} level and the SaTScan elimination reduces this to nine clusters of which only one appears to be almost stable as the flexibility is increased.

6. DISCUSSION

As more and more computational power becomes available it becomes feasible to run larger computations. This is most likely the reason why Geographic Information Systems (GIS) and analysis of geographic data are increasingly used. Following the increase in computer power there is a trend focusing on gathering more detailed datasets to be analyzed. When applying methods that originally were tested for use on more simple datasets one must be aware that some possible artifacts may come out.

It is well defined how to find the most likely cluster, but in cases with small populations or few observations on each geographical location the size is highly dependent on the outcome of the underlying process at the locations that are in the other rim and just outside the most likely cluster. Hence, the size is stochastic and this should be kept in mind when interpreting the most likely (and other) clusters. The effect may be more pronounced when looking at secondary clusters with the SaTScan elimination scheme. As long as the secondary clusters are found far away from more significant clusters it seems to be reasonable, but the elimination fails to find the true underlying clusters if the shape of those clusters are not in the chosen set of clusters; and also if the chosen set of shapes forces a cluster to cover more than the true underlying cluster and hence eliminates the neighboring underlying clusters. These effects become most pronounced when high- and low-risk clusters are geographically close under the given set of clusters.

One way to circumvent some of these problems is to create a more flexible set of clusters, and this work suggests creating clusters built up from confocal ellipses as they contain the set of concentric circles as a true subset. When a large proportion of the locations are included as secondary foci this gives rise to a very flexible set of clusters. We find that this flexibility is needed to avoid the problems with underlying clusters having odd shapes lying relatively close. The problem with such a flexible set is how to present the acquired information. In this study, 243,244 clusters were significant at a 10^{-4} level, they were reduced to 9 when the SaTScan elimination scheme was applied. Those 9 clusters make it hard to judge where the underlying problems are. This information, which must be the most important, is more easily accessible when all the highly significant clusters are represented. The disadvantage is that the borders of the areas of interest become less well defined. One could choose to select those locations that are part of more than a certain proportion of the highly significant clusters (in the present representation this corresponds to only selecting those locations having a bar longer than a selected threshold). Again one should keep in mind that the underlying process is stochastic and hence all defined borders of clusters should be interpreted with some flexibility.

In the simulation study it was shown that using elliptic scanning windows was as good or better than using circular scanning windows. In fact the error rate was halved in a large range of settings. This study was made using a true cluster with a circular shape it is expected that the benefit of using elliptic scanning windows will increase for other shapes of the underlying cluster.

When evaluating the additional flexibility provided by elliptic clusters one should keep in mind that rotating and scaling the coordinate system can transform an elliptic cluster into a circular cluster. Hence one should use an evaluation that is quasi-stable for smooth changes in the flexibility of the set of clusters under inspection. The previous section has exemplified that this is not the case then using the elimination scheme implemented in SaTScan, that is, nonoverlapping clusters.

The way chosen to evaluate and present the information may not be the best in all circumstances. For a given case one should try to present the results of the spatial scan statistic in several ways to get a better understanding of the underlying process. This could also include the use of the likelihood ratios to calculate a weighted proportion of clusters

that a given location belongs to.

This work is based on an all new implementation of the algorithms. It is written in C++ and OpenMP has been used to make a parallel version so that a multiprocessor computer can be used to speed up the calculations. Up to 20 CPUs on a SunFire 15K have been used in the calculations and more could easily have been used. The scalability is good as all the Monte Carlo simulations are independent and this is the most time consuming part of the calculations.

In conclusion, we have shown that the set of confocal ellipses creates a smooth expansion (with discrete steps due to the point process) of concentric circles. The presented simulation study shows that the set of confocal ellipses has the same or lower error rate when compared with circular clusters. Furthermore, we have provided examples of the problems that can occur when the SaTScan elimination scheme is applied, in particular rotations, scalings and clusters in close proximity. A new way to present the information and other alternatives have been suggested. It is found to be important to use more than one way to illustrate the information to avoid possible pitfalls.

ACKNOWLEDGMENTS

We acknowledge the Danish Center for Scientific Computing (DCSC) for the support.

[Received May 2004. Revised December 2005.]

REFERENCES

- Boscoe, F. P., McLaughlin, C., Schymura, M. J., and Kielb, C. L. (2003), "Visualization of the Spatial Scan Statistic Using Nested Circles," *Health and Place*, 9, 273–277.
- Enemark, H. L., Ahrens, P., Juel, C. D., Petersen, E., Petersen, R. F., Andersen, J. S., Lind, P., and Thamsborg, S. M. (2002), "Molecular Characterization of Danish *Cryptosporidium parvum* Isolates," *Parasitology*, 125, 331–341.
- Gangnon, R. E., and Clayton, M. K. (2001), "A Weighted Average Likelihood Ratio Test for Spatial Clustering of Disease," *Statistics in Medicine*, 20, 2977–2987.
- Hoar, B. R., Chomel, B. B., Rolfé, D. L., Chang, C. C., Fritz, C. L., Sacks, B. N., and Carpenter, T. E. (2003), "Spatial Analysis of *Yersinia pestis* and *Bartonella vinsonii* subsp. *berkhoffii* Seroprevalence in California Coyotes (*Canis latrans*)," *Preventive Veterinary Medicine*, 56, 299–311.
- Kulldorff, M. (1997), "A Spatial Scan Statistic," *Communications in Statistics: Theory and Methods*, 26, 1481–1496.
- (1999), "Spatial Scan Statistics: Models, Calculations and Applications," in *Recent Advances on Scan Statistics and Applications*, Boston: Birkhauser, pp. 303–322.
- Patrick, M. E., Christiansen, L. E., Wainø, M., Ethelberg, S., Madsen, H., and Wegener, H. C. (2004), "Effects of Climate on Incidence of *Campylobacter* spp. in Humans and Prevalence in Broiler Flocks in Denmark," *Applied and Environmental Microbiology*, 70, 7474–7480.
- Roche, L. M., Skinner, R., and Weinstein, R. B. (2002), "Use of a Geographic Information System to Identify and Characterize Areas with High Proportions of Distant Stage Breast Cancer," *Journal of Public Health Management and Practice*, 8, 26–32.
- Sauders, B. D., Fortes, E. D., Morse, D. L., Dumas, N., Kiehlbauch, J. A., Schukken, Y., Hibbs, J. R., and Wiedmann, M. (2003), "Molecular Subtyping to Detect Human Listeriosis Clusters," *Emerging Infectious Diseases*, 9, 672–680.