SPATIAL SCAN STATISTIC: SELECTING CLUSTERS AND GENERATING ELLIPTIC CLUSTERS

L.E. Christiansen[1,2] and J.S. Andersen[2]

[1] Informatics and Mathematical Modelling, Technical University of Denmark, DTU Building 321, DK-2800 Kgs. Lyngby, Denmark; [2] Danish Institute for Food and Veterinary Research, Mørkhøj Bygade 19, DK-2860 Søborg, Denmark.

Summary

The spatial scan statistic is widely used to search for clusters. This paper shows that the usually applied elimination of overlapping clusters to find secondary clusters is sensitive to smooth changes in the shape of the clusters. We present an algorithm for generation of set of confocal elliptic clusters. In addition, we propose a new way to present the information in a given set of clusters based on the significance of the clusters.

Introduction

In epidemiology the spatial scan statistic as implemented in *SaTScan* (www.SaTScan.org) is widely used to search for clusters in spatial data. *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 - especially if a case with hundreds of spatial locations is considered. The question is how much information is lost during this reduction? It is a fact that not all naturally occurring phenomena are circular and hence there is a need for methods that can find clusters with other shapes.

Brief model description

We present a spatial scan statistic that generates a set of subsets of confocal ellipses, which includes the set of concentric circles used in *SaTScan* as a true subset. Set of confocal ellipses is constructed in a similar way to set of concentric ellipses[(2)]. At first all points are used to create concentric clusters; next pairs of two points are used as foci for ellipses, this is done by starting with one point as the primary focus and then using the nearest proportion (p_s) of points as the secondary focus. Additional points are included according to the sum of the distances to the two foci. A more thorough description will be published[(1)].

The null hypothesis is that the incidence rate is the same all over the study region and the alternative is that it is higher in a given subset. In order to test if the null hypothesis holds we use Monte Carlo simulations. The number of simulations depends on which confidence limit to be used. In the case of a 1% significance level and 9,999 simulations the most likely subset is significant if its likelihood ratio is among the 99 highest simulated maximum likelihood ratios.

If the null hypothesis falls for the most likely subset the question is how about all the other subsets? The subsets having one additional point or lacking one point when compared with the most likely subset will in many cases have likelihood ratios just below the most likely one, this is to remind us that the underlying cluster may be of a different size and shape. Furthermore, one should keep in mind that the dataset is a stochastic point process.

It is also of interest if there are other significant subsets representing totally different clusters.

In a case with many subsets having a likelihood ratio above the level of significance it becomes difficult to present the information. We propose that for each location the proportion of significant subsets that it belongs to is used as a measure (it could be weighted using the likelihood ratios). The algorithms are coded in C++, and OpenMP was used to parallelize the Monte Carlo simulations. The code will be available at request.

The case that is used for illustration in this work 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 time 23,279 broiler flock samples were collected; only 8,056 were included in the analysis; the remaining were excluded from the data set either because they were not from the first batch slaughtered (the risk increase for the following batches), were reported to come from an unknown house, or not located within Jutland and Funen (Containing most of the production and the remaining broiler flocks are located far from this area). In 3,080 of the flocks a *Campylobacter* infection was found.

Results

For the set of circles 21,324 out of the 107,953 different subsets including up to a proportion p_m=0.50 of the points were significant at a 10^{-4} level; based on 99,999 Monte Carlo simulations. When applying the elimination scheme implemented in SaTScan, removing overlapping clusters, only 9 clusters are left (*Figure. 1a*). The method we propose (*Figure. 1c*) agrees on the two most likely non-overlapping clusters, but shows that the points in the lower ranking non-overlapping secondary clusters (No. 3-9) seem to be of less interest even though they are significant.

When considering ellipses with p_s=0.20 and including up to half of the points the result found with non-overlapping clusters (*Figure. 1b*) seem to be inconsistent where as the proposed method (*Figure. 1d*) indeed is more stable.

Discussion

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, in particular at the locations in the outer 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.

We find that the proposed way to illustrate a set of clusters is a beneficial alternative, especially when the true underlying cluster is not covered by the chosen shape of clusters.

References

1. Christiansen LE, Andersen JS, Wegener HC and Madsen H (*submitted*) Spatial scan statistics using elliptic clusters. *Journal of Agricultural, Biological, and Environmental Statistics*
2. Kulldorff M (1997). A spatial scan statistic. *Commun. Stat.-Theor. M.* 26,1481-1496.

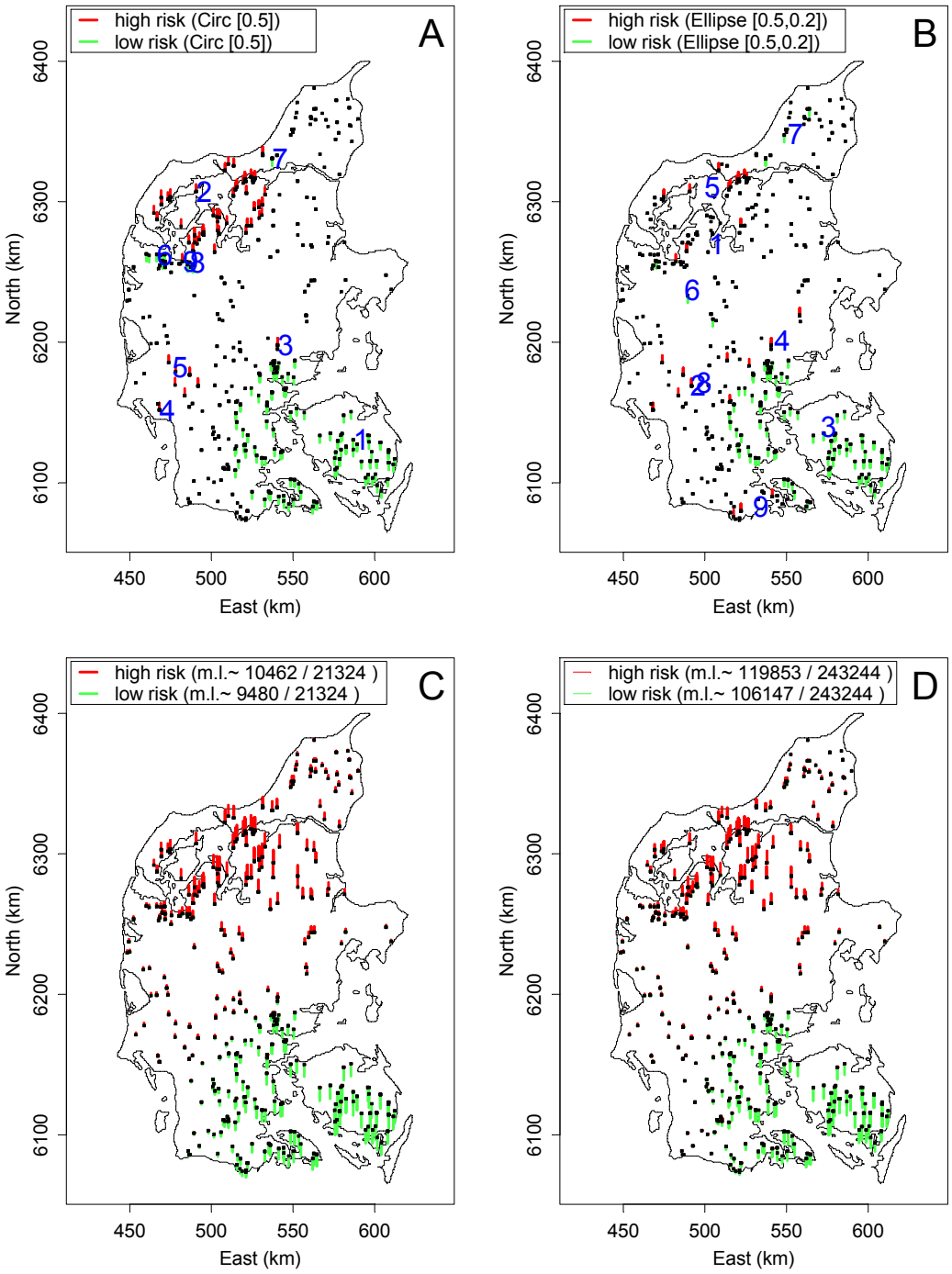

Figure 1. Clustering of *Campylobacter* in broilers. Black dots are locations of farms. The red and green bars show occurrences in high and low risk clusters, respectively. The maximum cluster size was 50% of the farms. Subfigure A and C present the result of circular clusters. Subfigure B and D present the results of elliptic clusters with the nearest 20% as the secondary focus. In Subfigure A and B search for secondary clusters only include farms not belonging to detected higher ranking clusters (as in SaTScan). The numbers (1-9) in the plot are the ranks of the clusters. In Subfigure C and D the length of the bars indicates the number of occurrences in all the significant clusters; the highest occurrences in high and low risk clusters are stated in the legend.