SatScan

- The general statistical theory behind the spatial and space-time scan statistic is described in detail by Kulldorff (1997). Here we give a brief non-mathematical description of the particular models that can be analyzed using the SaTScan software. Special cases have been described by Kulldorff et al. (1995), Hjalmars et al. (1996) and Kulldorff et al. (1997) in connection with specific applications.
- Bernoulli and Poisson Models

The SaTScan software can analyse two different probabilistic models, based on the Bernoulli and Poisson distributions respectively.

With the Bernoulli model, there are cases and non-cases as a 0/1 variable. These may represent people with or without a disease, or people with different types of disease. They may reflect cases and controls from a larger population, or they may together constitute the population as a whole. Whatever the situation may be, they will be denoted as cases and controls throughout the help files, and their total will be denoted as the population.

- With the Poisson model, the number of cases in each census area is assumed to be Poisson distributed. Under the null hypothesis, and when there are no covariates, the expected number of cases in each area is proportional to its population size, or to the person-years in that area. When there are covariates, the covariate adjusted expected number of cases is used.
- With either model, the scan statistic adjusts for the uneven population density present in almost all populations, and the analysis is conditioned on the total number of cases observed.

Data Requirements

For the Poisson model, it is necessary to have case and population counts for a set of census areas, as well as the geographical coordinates for each of those areas. For the Bernoulli model the population counts are replaced by the number of controls. Separate census areas may be specified for individuals or data may be aggregated for states, provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc. To do a space-time analysis, it is also necessary to have a time related to each case, and with the Bernoulli model, for each control as well.

The Spatial Scan Statistic

The spatial scan statistic imposes a circular window on the map. The window is in turn centered around each of several possible centroids positioned throughout the study region. For each centroid, the radius of the window varies continuously in size from zero to some upper limit

. In this way, the circular window is flexible both in location and size. In total, the method creates an infinite number of distinct geographical circles, with different sets of neighboring census areas within them, and each being a possible candidate for a cluster.

The set of centroids used is defined either in a special grid file, or they are taken to be identical to the different census locations as specified in the coordinates file. The latter option ensures that each census area is a potential cluster in itself.

The Space-Time Scan Statistic

The space-time scan statistic is defined by a cylindrical window with a circular geographic base and with height corresponding to time. The base is defined exactly as for the purely spatial scan statistic, while the height reflects the time period of potential clusters. The cylindrical window is then moved in space and time, so that for each possible geographical location and size, it also visits each possible time period. In effect, we obtain an infinite number of overlapping cylinders of different size and shape, jointly covering the entire study region, where each cylinder reflects a possible cluster.

The Temporal Scan Statistic

The temporal scan has a window that moves in one dimension (time), defined in the same way as the height of the cylinder used by the space-time scan statistic. This means that it is flexible in both location and size, covering anything from the length of a time interval to the maximum temporal cluster size as specified on the Scanning Window tab.

Covariates

With the Poisson model it is possible to adjust for any number of categorical covariates. If the disease rate varies with for example age, and if the age distribution is different in different areas, then there is geographical clustering of the disease simply due to the age covariate. By specifying categorical covariates, the SaTScan program will search for clusters above and beyond that which is expected due to the covariates. For the Poisson model, the expected number of cases in each area under the null-hypothesis, is then calculated based on the covariates, using indirect standardization. When more than one covariate is specified, each one is adjusted for as well as the interaction terms between them.

The user may also himself or herself calculate the covariate adjusted expected number of cases. These should then replace the raw population numbers in the population file, while not including the covariates themselves. In this way, it is also possible to adjust for continuous covariates.

Likelihood Ratio Test

For each location and size of the scanning window, the alternative hypothesis is that there is an elevated rate within the window as compared to outside. Under the Poisson assumption, the likelihood function for a specific window is then proportional to

$$(n/\mu)^{n}([N-n]/[N-n])^{(N-n)}I()$$

where N is the total number of cases over the whole area, n is the number of cases within the window, and μ: is the covariate adjusted expected number of cases within the window under the null-hypothesis.

For the Bernoulli model the likelihood function is instead

$$(n/m)^{n}(1-n/m)^{(m-n)}((N-n)/(M-m))^{(N-n)}(1-((N-n)/(M-m)))^{(M-m)-(N-n)}I()$$

I() is an indicator function. When SaTScan is set to scan only for clusters with high rates, than I() is equal to 1 when the window has more cases than expected under the null-hypothesis and 0 otherwise. The opposite is true when SaTScan is set to scan only for clusters with low rates. When the program scans for clusters with either high or low rates, than I()=1 for all windows.

The likelihood function is maximized over all windows, identifying the window that constitutes the most likely cluster. This is the cluster that is least likely to have occurred by chance. The likelihood ratio for this window is noted and constitutes the maximum likelihood ratio test statistic. Its distribution under the null-hypothesis and its corresponding p-value is obtained by repeating the same analytic exercise on a large number random of replications of the data set generated under the null hypothesis, in a Monte Carlo simulation.

Secondary Clusters

In addition to the most likely cluster, the method also identifies secondary clusters in the data set, and orders them according to their likelihood ratio. There will always be sets of counties that overlap in part with the most likely cluster and that have almost as high likelihood value, since adding or subtracting a few census areas often does not change the likelihood greatly. The SaTScan program does not report clusters of this type since most of them provide little additional information, but their existence means that while it is possible to pinpoint the general location of a cluster, its exact boundaries must remain uncertain.

There may also be secondary clusters that do not overlap the most likely cluster. The SaTScan software reports such clusters if, (1) it has the highest likelihood function for a particular centroid, and (2) it does not overlap with the most likely cluster nor with a previously reported secondary cluster of higher likelihood.

Population Data

Population data need not be specified continuously over time, but only at one or more specific census times. For times in between, the SaTScan program makes a linear interpolation based on the population at the census times immediately proceeding and immediately following. For times before the first census time, the population size is set equal to the its size at that census time, and for times after the last census time, the equivalent is done.

To get the population size for a given census area and time period, the census area population size, as defined above, is integrated over the time period in question.