

---

# **SaTScan<sup>TM</sup> User Guide**

**for version 9.2**

**By Martin Kulldorff**

**July, 2013**

**<http://www.satscan.org/>**

# Contents

<b>Introduction .....</b>	<b>4</b>
The SaTScan Software .....	4
Download and Installation.....	5
Test Run .....	5
Help System .....	6
Sample Data Sets.....	6
<b>Statistical Methodology.....</b>	<b>10</b>
Spatial, Temporal and Space-Time Scan Statistics .....	11
Bernoulli Model .....	12
Discrete Poisson Model.....	13
Space-Time Permutation Model.....	13
Multinomial Model .....	14
Ordinal Model .....	15
Exponential Model .....	15
Normal Model .....	16
Continuous Poisson Model.....	17
Probability Model Comparison .....	18
Likelihood Ratio Test.....	19
Secondary Clusters .....	21
Adjusting for More Likely Clusters.....	21
Covariate Adjustments .....	22
Spatial and Temporal Adjustments.....	25
Missing Data .....	27
Multivariate Scan with Multiple Data Sets.....	28
<b>Comparison with Other Methods.....</b>	<b>30</b>
Scan Statistics.....	30
Spatial and Space-Time Clustering .....	30
<b>Input Data .....</b>	<b>32</b>
Data Requirements .....	32
Case File.....	33
Control File .....	34
Population File .....	34
Coordinates File .....	35
Grid File .....	36
Non-Euclidian Neighbors File.....	37
Meta Location File .....	37
Max Circle Size File.....	38
Adjustments File.....	38
Alternative Hypothesis File.....	39
SaTScan Import Wizard .....	40
SaTScan ASCII File Format.....	41
<b>Basic SaTScan Features .....</b>	<b>44</b>
Input Tab .....	44
Analysis Tab.....	47
Output Tab .....	51
<b>Advanced Features .....</b>	<b>54</b>
Multiple Data Sets Tab.....	54
Data Checking Tab.....	55
Spatial Neighbors Tab .....	56
Spatial Window Tab.....	58
Temporal Window Tab .....	61

Spatial and Temporal Adjustments Tab .....	63
Inference Tab .....	65
Power Estimation Tab .....	68
Spatial Output Tab .....	69
Other Output Tab .....	72
<b>Running SaTScan .....</b>	<b>74</b>
Specifying Analysis and Data Options .....	74
Launching the Analysis .....	74
Status Messages.....	75
Warnings and Errors.....	75
Saving Analysis Parameters .....	76
Parallel Processors .....	77
Batch Mode .....	77
Computing Time.....	78
Memory Requirements .....	79
<b>Results of Analysis.....</b>	<b>82</b>
Standard Text-Based Results File (*.out.*) .....	82
Shapefile Geographical Output (*.shp and *.shx) .....	84
Cluster Information File (*.col.*) .....	84
Stratified Cluster Information File (*.sci.*) .....	86
Location Information File (*.gis.*) .....	86
Risk Estimates for Each Location File (*.rr.*) .....	87
Simulated Log Likelihood Ratios File (*.llr.*) .....	87
<b>Miscellaneous .....</b>	<b>88</b>
New Versions .....	88
Analysis History File .....	88
Random Number Generator .....	88
Contact Us.....	89
Acknowledgements .....	89
<b>Frequently Asked Questions.....</b>	<b>91</b>
Input Data.....	91
Analysis.....	92
Results.....	93
Interpretation.....	93
Operating Systems.....	95
<b>SaTScan Bibliography.....</b>	<b>96</b>
Suggested Citations .....	96
SaTScan Methodology Papers.....	97
Selected SaTScan Applications by Field of Study .....	101
Other References Mentioned in the User Guide .....	116

# Introduction

---

## The SaTScan Software

### *Purpose*

SaTScan is a free software that analyzes spatial, temporal and space-time data using the spatial, temporal, or space-time scan statistics. It is designed for any of the following interrelated purposes:

- Perform geographical surveillance of disease, to detect spatial or space-time disease clusters, and to see if they are statistically significant.
- Test whether a disease is randomly distributed over space, over time or over space and time.
- Evaluate the statistical significance of disease cluster alarms.
- Perform prospective real-time or time-periodic disease surveillance for the early detection of disease outbreaks.

The software may also be used for similar problems in other fields such as archaeology, astronomy, botany, criminology, ecology, economics, engineering, forestry, genetics, geography, geology, history, neurology or zoology.

### *Data Types and Methods*

SaTScan can be used for discrete as well as continuous scan statistics. For discrete scan statistics the geographical locations where data are observed are non-random and fixed by the user. These locations may be the actual locations of the observations, such as houses, schools or ant nests, or it could be a central location representing a larger area, such as the geographical or population weighted centroids of postal areas, counties or provinces. For continuous scan statistics, the locations of the observations are random and can occur anywhere within a predefined study area defined by the user, such as a rectangle.

For discrete scan statistics, SaTScan uses either a discrete Poisson-based model, where the number of events in a geographical location is Poisson-distributed, according to a known underlying population at risk; a Bernoulli model, with 0/1 event data such as cases and controls; a space-time permutation model, using only case data; a multinomial model for categorical data; an ordinal model, for ordered categorical data; an exponential model for survival time data with or without censored variables; a normal model for other types of continuous data; or a spatial variation in temporal trends model, looking for geographical areas with unusually high or low temporal trends. A common feature of all these discrete scan statistics is that the geographical locations where data can be observed are non-random and fixed by the user.

For the discrete scan statistics, the data may be either aggregated at the census tract, zip code, county or other geographical level, or there may be unique coordinates for each observation. SaTScan adjusts for the underlying spatial inhomogeneity of a background population. It can also adjust for any number of categorical covariates provided by the user, as well as for temporal trends, known space-time clusters and missing data. It is possible to scan multiple data sets simultaneously to look for clusters that occur in one or more of them.

For continuous scan statistics, SaTScan uses a continuous Poisson model.

## Developers and Funders

The SaTScan™ software was developed by Martin Kulldorff together with Information Management Services Inc. Financial support for SaTScan has been received from the following institutions:

- National Cancer Institute, Division of Cancer Prevention, Biometry Branch [v1.0, 2.0, 2.1]
- National Cancer Institute, Division of Cancer Control and Population Sciences, Statistical Research and Applications Branch [v3.0 (part), v6.1 (part), 8.0 (part), v9.0 (part)]
- Alfred P. Sloan Foundation, through a grant to the New York Academy of Medicine (Farzad Mostashari, PI) [v3.0 (part), 3.1, 4.0, 5.0, 5.1]
- Centers for Disease Control and Prevention, through Association of American Medical Colleges Cooperative Agreement award number MM-0870 [v6.0, 6.1 (part)].
- National Institute of Child Health and Development, through grant #RO1HD048852 [7.0, 8.0, 9.0 (part)]
- National Cancer Institute, Division of Cancer Epidemiology and Genetics [v9.0 (part), v9.1]
- National Institute of General Medical Sciences, through a Modeling Infectious Disease Agent Studies grant #U01GM076672 [v9.0 (part)]

Their financial support is greatly appreciated. The contents of SaTScan are the responsibility of the developer and do not necessarily reflect the official views of the funders.

**Related Topics:** *Statistical Methodology, SaTScan Bibliography*

---

## Download and Installation


To install SaTScan, go to the SaTScan Web site at: <http://www.satscan.org/> and select the SaTScan download link. After downloading the SaTScan installation executable to your PC, click on its icon and install the software by following the step-wise instructions.

**Related Topics:** *New Versions.*

---

## Test Run

Before using your own data, we recommend trying one of the sample data sets provided with the software. Use these to get an idea of how to run SaTScan. To perform a test run:

1. Click on the SaTScan application icon.
2. Click on ‘Open Saved Session’.
3. Select one of the parameter files, for example ‘nm.prm’ (Poisson model), ‘NHumberside.prm’ (Bernoulli model) or ‘NYCfever.prm’ (space-time permutation model).
4. Click on ‘Open’.
5. Click on the Execute  button. A new window will open with the program running in the top section and a Warnings/Errors section below. When the program finishes running the results will be displayed.

**Note:** The sample files should not produce warnings or errors.

**Related Topics:** *Sample Data Sets.*

---

## Help System

The SaTScan help system consists of four parts:

- i. SaTScan User Guide in PDF format, located in the same folder as the SaTScan executable. It can also be obtained from the SaTScan web site ([www.satscan.org/techdoc.html](http://www.satscan.org/techdoc.html)) or directly within the SaTScan software by selecting Help > User Guide. You may print this as a single document for easy reference.
- ii. SaTScan help entries, extracted from the User Guide. The complete set of entries can be found within the SaTScan software by typing F1 or by selecting Help > Help Content. The system can be searched by clicking the magnifying glass. Many individual entries can also be reached directly by clicking on any sub-title seen on the input tabs.
- iii. Methodological papers describe the details about the statistical methods available in the SaTScan software. These papers are listed in the SaTScan bibliography, which can be found both at the end of the User Guide and on the web (<http://www.satscan.org/references.html>). The bibliography also contains a large number of papers that have applied different SaTScan features for a wide range of different types of data. These can serve as inspiration for how SaTScan can be used for different types of scientific and public health problems.
- iv. Sample data sets are provided for each of the SaTScan probability models. Described below, they make it easy to familiarize oneself with the software.

---

## Sample Data Sets

Six different sample data sets are provided with the software. They are automatically downloaded to your computer together with the software itself. Other sample data sets are available at <http://www.satscan.org/datasets/>.

### ***Discrete Poisson Model, Space-Time and Spatial Variation in Temporal Trends: Brain Cancer Incidence in New Mexico***

**Case file:** nm.cas

**Format:** <county> <cases=1> <year> <age group> <sex>

**Population file:** nm.pop

**Format:** <county> <year> <population> <age group> <sex>

**Coordinates file:** nm.geo

**Format:** <county> <x-coordinate> <y-coordinate>

**Study period:** 1973-1991

**Aggregation:** 32 counties

**Precision of case times:** Years

**Coordinates:** Cartesian

**Covariate #1, age groups:** 1 = 0-4 years, 2 = 5-9 years, ... 18 = 85+ years

**Covariate #2, gender:** 1 = male, 2 = female

**Population years:** 1973, 1982, 1991

**Data source:** New Mexico SEER Tumor Registry

This is a condensed version of a more complete data set with the population given for each year from 1973 to 1991, and with ethnicity as a third covariate. The complete data set can be found at: <http://www.satscan.org/datasets/>

***Bernoulli Model, Purely Spatial: Childhood Leukemia and Lymphoma Incidence in North Humberside***

**Case file:** NHumberside.cas

**Format:** <location id> <# cases>

**Control file:** Nhumberside.ctl

**Format:** <location id> <# controls>

**Coordinates file:** Nhumberside.geo

**Format:** <location id> <x-coordinate> <y-coordinate>

**Study period:** 1974-1986

**Controls:** Randomly selected from the birth registry

**Aggregation:** 191 Postal Codes (most with only a single individual)

**Precision of case and control times:** None

**Coordinates:** Cartesian

**Covariates:** None

**Data source:** Drs. Ray Cartwright and Freda Alexander. Published by J. Cuzick and R. Edwards, Journal of the Royal Statistical society, B:52 73-104, 1990

***Space-Time Permutation Model: Hospital Emergency Room Admissions Due to Fever at New York City Hospitals***

**Case file:** NYCfever.cas

**Format:** <zip> <#cases=1> <date>

**Coordinates file:** NYCfever.geo

**Format:** <zip> <latitude> <longitude>

**Study period:** Nov 1, 2001 – Nov 24, 2001

**Aggregation:** Zip code areas

**Precision of case times:** Days

**Coordinates:** Latitude/Longitude

**Covariates:** None

**Data source:** New York City Department of Health

***Multinomial and Ordinal Model, Purely Spatial: Education Attainment Levels in Maryland***

**Case file:** MarylandEducation.cas

**Format:** <county> <# individuals> <category #>

**Coordinates file:** MarylandEducation.geo

**Format:** <county> <latitude> <longitude>

**Study period:** 2000

**Aggregation:** 24 Counties and County Equivalents

**Precision of case times:** None

**Coordinates:** Latitude / Longitude

**Covariates:** None

**Categories:** 1 = Less than 9<sup>th</sup> grade  
2 = 9<sup>th</sup> to 12<sup>th</sup> grade, but no high school diploma  
3 = High school diploma, but no bachelor degree  
4 = Bachelor or higher degree

**Data source:** United States Census Bureau: Information about education comes from the long Census 2000 form, filled in by about 1/6 households.

**Note:** Only people age 25 and above are included in the data. For each county, the census provides information about the percent of people with different levels of formal education. The number of individuals reporting different education levels in each county was estimated as this percentage times the total population age 25+ divided by six to reflect the 1/6 sampling fraction for the long census form.

***Exponential Model, Space-Time : Artificially Created Survival Data***

**Case file:** SurvivalFake.cas

**Format:** <location id> <# individuals> <time of diagnosis> <survival time> <censored>

**Coordinates file:** SurvivalFake.geo

**Format:** <location id> <x-coordinate> <y-coordinate>

**Study period:** 2000-2005

**Aggregation:** 5 Locations

**Precision of times of diagnosis:** Year

**Precision of survival/censoring times:** Day

**Coordinates:** Cartesian

**Covariates:** None



**Data source:** Artificially created data.

***Normal Model, Purely Spatial : Artificially Created Continuous Data***

**Case file:** NormalFake.cas

**Format:** <location id> <# individuals> <weight increase>

**Coordinates file:** NormalFake.geo

**Format:** <location id> <x-coordinate> <y-coordinate>

**Study period:** 2006

**Aggregation:** 26 Locations

**Coordinates:** Cartesian

**Covariates:** None

**Data source:** Artificially created data.

**Related Topics:** *Test Run, Input Data.*

# Statistical Methodology

For all discrete spatial and space-time analyses, the user must provide data containing the spatial coordinates of a set of locations (coordinates file). For each location, the data must furthermore contain information about the number of cases at that location (case file). For temporal and space-time analyses, the number of cases must be stratified by time, e.g. the time of diagnosis. Depending on the type of analysis, other information about cases such as age, gender, weight, length of survival and/or cancer stage may also be provided. For the Bernoulli model, it is also necessary to specify the number of controls at each location (control file). For the discrete Poisson model, the user must specify a population size for each location (population file). The population may vary over time.

Scan statistics are used to detect and evaluate clusters of cases in either a purely temporal, purely spatial or space-time setting. This is done by gradually scanning a window across time and/or space, noting the number of observed and expected observations inside the window at each location. In the SaTScan software, the scanning window is an interval (in time), a circle or an ellipse (in space) or a cylinder with a circular or elliptic base (in space-time). It is also possible to specify your own non-Euclidian distance structure in a special file. Multiple different window sizes are used. The window with the maximum likelihood is the most likely cluster, that is, the cluster least likely to be due to chance. A p-value is assigned to this cluster.

Scan statistics use a different probability model depending on the nature of the data. A Bernoulli, discrete Poisson or space-time permutation model is used for count data such as the number of people with asthma; a multinomial model is used for categorical data such as cancer histology; an ordinal model for ordered categorical data such as cancer stage; an exponential model for survival time data with or without censoring; and a normal model for other continuous data such as birth weight or blood lead levels. The general statistical theory behind the spatial and space-time scan statistics used in the SaTScan software is described in detail by Kulldorff (1997)<sup>1</sup> for the Bernoulli, discrete Poisson and continuous Poisson models; by Kulldorff et al. (2005)<sup>5</sup> for the space-time permutation model; by Jung et al. (2008)<sup>6</sup> for the multinomial model; by Jung et al. (2007)<sup>7</sup> for the ordinal model; by Huang et al. (2006)<sup>8</sup> for the exponential model, by Kulldorff et al. (2009)<sup>9</sup> for the normal model and by Huang et al. (2009)<sup>10</sup> for the normal model with weights. Please read these papers for a detailed description of each model. Here we only give a brief non-mathematical description.

For all discrete probability models, the scan statistic adjusts for the uneven geographical density of a background population. For all models, the analyses are conditioned on the total number of cases observed.

**Related Topics:** *The SaTScan Software, Basic SaTScan Features, Advanced Features, Analysis Tab, Methodological Papers.*

---

# Spatial, Temporal and Space-Time Scan Statistics

## Spatial Scan Statistic

The standard purely spatial scan statistic imposes a circular window on the map. The window is in turn centered on each of several possible grid points positioned throughout the study region. For each grid point, the radius of the window varies continuously in size from zero to some upper limit specified by the user. 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 data locations within them. Each circle is a possible candidate cluster.

The user defines the set of grid points used through a grid file. If no grid file is specified, the grid points are set to be identical to the coordinates of the location IDs defined in the coordinates file. The latter option ensures that each data location is a potential cluster in itself, and it is the recommended option for most types of analyses.

As an alternative to the circle, it is also possible to use an elliptic window shape, in which case a set of ellipses with different shapes and angles are used as the scanning window together with the circle. This provides slightly higher power for true clusters that are long and narrow in shape, and slightly lower power for circular and other very compact clusters.

It is also possible to define your own non-Euclidian distance metric using a special neighbors file.

**Related Topics:** *Analysis Tab, Coordinates File, Elliptic Scanning Window, Grid File, Maximum Spatial Cluster Size, Spatial Window Tab.*

## Space-Time Scan Statistic

The space-time scan statistic is defined by a cylindrical window with a circular (or elliptic) 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 space-time scan statistic may be used for either a single retrospective analysis, using historic data, or for time-periodic prospective surveillance, where the analysis is repeated for example every day, week, month or year.

**Related Topics:** *Analysis Tab, Spatial Window Tab, Temporal Window Tab.*

## Temporal Scan Statistic

The temporal scan statistic uses 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 start and end date. The maximum temporal length is specified on the Temporal Window Tab.

**Related Topics:** *Analysis Tab, Temporal Window Tab, Space-Time Scan Statistic.*

### **Spatial Variation in Temporal Trends Scan Statistic**

When the scan statistic is used to evaluate the spatial variation in temporal trends, the scanning window is purely spatial in nature. The temporal trend is then calculated inside as well as outside the scanning window, for each location and size of that window. The null hypothesis is that the trends are the same, while the alternative is that they are different. Based on these hypotheses, a likelihood is calculated, which is higher the more unlikely it is that the difference in trends is due to chance. The most likely cluster is the cluster for which the temporal trend inside the window is least likely to be the same as the temporal trend outside the cluster. This could be because of various reasons. For example, if the temporal trend inside the cluster is higher, it could be because all areas has the same incidence rate of a disease at the beginning of the time period, but the cluster area has a higher rate at the end of the time period. It could also be because the cluster area has a lower incidence rate at the beginning of the time period, after which it ‘catches up’ with the rest so that the rate is about the same at the end of the time period. Hence, a statistically significant cluster in the spatial variation in temporal trend analysis does not necessarily mean that the overall rate of disease is higher or lower in the cluster.

The spatial variation in temporal trends scan statistic can only be run with the discrete Poisson probability model. For it to work, it is important that the total study period length is evenly divisible by the length of the time interval aggregation, so that all time intervals have the same number of years, if it is specified in years, the same number of months if it is specified in months or the same number of days if it is specified in days.

**Related Topics:** *Analysis Tab, Space-Time Scan Statistic.*

---

## **Bernoulli Model**

With the Bernoulli model<sup>1,2</sup>, there are cases and non-cases represented by a 0/1 variable. These variables may represent people with or without a disease, or people with different types of disease such as early and late stage breast cancer. 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, these variables will be denoted as cases and controls throughout the user guide, and their total will be denoted as the population. Bernoulli data can be analyzed with the purely temporal, the purely spatial or the space-time scan statistics.

*Example:* For the Bernoulli model, cases may be newborns with a certain birth defect while controls are all newborns without that birth defect.

The Bernoulli model requires information about the location of a set of cases and controls, provided to SaTScan using the case, control and coordinates files. Separate locations may be specified for each case and each control, or the data may be aggregated for states, provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc, with multiple cases and controls at each data location. To do a temporal or space-time analysis, it is necessary to have a time for each case and each control as well.

**Related Topics:** *Analysis Tab, Case File, Control File, Coordinates File, Likelihood Ratio Test, Methodological Papers, Probability Model Comparison.*

---

## Discrete Poisson Model

With the discrete Poisson model<sup>1</sup>, the number of cases in each location is 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. Poisson data can be analyzed with the purely temporal, the purely spatial, the space-time and the spatial variation in temporal trends scan statistics.

*Example:* For the discrete Poisson model, cases may be stroke occurrences while the population is the combined number of person-years lived, calculated as 1 for someone living in the area for the whole time period and  $\frac{1}{2}$  for someone dying or moving away in the middle of the time period.

The discrete Poisson model requires case and population counts for a set of data locations such as counties, parishes, census tracts or zip code areas, as well as the geographical coordinates for each of those locations. These need to be provided to SaTScan using the case, population and coordinates files.

The population data need not be specified continuously over time, but only at one or more specific ‘census times’. For times in between, SaTScan does a linear interpolation based on the population at the census times immediately preceding and immediately following. For times before the first census time, the population size is set equal to the population size at that first census time, and for times after the last census time, the population is set equal to the population size at that last census time. To get the population size for a given location and time period, the population size, as defined above, is integrated over the time period in question.

**Related Topics:** *Analysis Tab, Case File, Continuous Poisson Model, Coordinates File, Likelihood Ratio Test, Methodological Papers, Population File, Probability Model Comparison.*

---

## Space-Time Permutation Model

The space-time permutation model<sup>5</sup> requires only case data, with information about the spatial location and time for each case, with no information needed about controls or a background population at risk. The number of observed cases in a cluster is compared to what would have been expected if the spatial and temporal locations of all cases were independent of each other so that there is no space-time interaction. That is, there is a cluster in a geographical area if, during a specific time period, that area has a higher proportion of its cases in that time period compared to the remaining geographical areas. This means that if, during a specific week, all geographical areas have twice the number of cases than normal, none of these areas constitute a cluster. On the other hand, if during that week, one geographical area has twice the number of cases compared to normal while other areas have a normal amount of cases, then there will be a cluster in that first area. The space-time permutation model automatically adjusts for both purely spatial and purely temporal clusters. Hence there are no purely temporal or purely spatial versions of this model.

*Example:* In the space-time permutation model, cases may be daily occurrences of ambulance dispatches to stroke patients.

It is important to realize that space-time permutation clusters may be due either to an increased risk of disease, or to different geographical population distribution at different times, where for example the population in some areas grows faster than in others. This is typically not a problem if the total study

period is less than a year. However, the user is advised to be very careful when using this method for data spanning several years. If the background population increases or decreases faster in some areas than in others, there is risk for population shift bias, which may produce biased p-values when the study period is longer than a few years. For example, if a new large neighborhood is developed, there will be an increase in cases there simply because the population increases, and using only case data, the space-time permutation model cannot distinguish an increase due to a local population increase versus an increase in the disease risk. As with all space-time interaction methods, this is mainly a concern when the study period is longer than a few years<sup>180,182</sup>. If the population increase (or decrease) is the same across the study region, that is okay, and will not lead to biased results.

**Related Topics:** *Analysis Tab, Case File, Coordinates File, Likelihood Ratio Test, Methodological Papers, Probability Model Comparison.*

---

## Multinomial Model

With the multinomial model<sup>6</sup>, each observation is a case, and each case belongs to one of several categories. The multinomial scan statistic evaluates whether there are any clusters where the distribution of cases is different from the rest of the study region. For example, there may be a higher proportion of cases of types 1 and 2 and a lower proportion of cases of type 3 while the proportion of cases of type 4 is about the same as outside the cluster. If there are only two categories, the ordinal model is identical to the Bernoulli model, where one category represents the cases and the other category represents the controls. The cases in the multinomial model may be a sample from a larger population or they may constitute a complete set of observations. Multinomial data can be analyzed with the purely temporal, the purely spatial or the space-time scan statistics.

*Example:* For the multinomial model, the data may consist of everyone diagnosed with meningitis, with five different categories representing five different clonal complexes of the disease<sup>6</sup>. The multinomial scan statistic will simultaneously look for high or low clusters of any of the clonal complexes, or a group of them, adjusting for the overall geographical distribution of the disease. The multiple comparisons inherent in the many categories used are accounted for when calculating the p-values.

The multinomial model requires information about the location of each case in each category. A unique location may be specified for each case, or the data may be aggregated for states, provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc, with multiple cases in the same location. To do a temporal or space-time analysis, it is necessary to have a time for each case as well.

With the multinomial model it is not necessary to specify a search for high or low clusters, since there is no hierarchy among the categories, but in the output it is shown what categories are more prominent inside the cluster. The order or indexing of the categories does not affect the analysis in terms of the clusters found, but it may influence the randomization used to calculate the p-values.

**Related Topics:** *Analysis Tab, Case File, Coordinates File, Likelihood Ratio Test, Methodological Papers, Probability Model Comparison.*

---

## Ordinal Model

With the ordinal model<sup>7</sup>, each observation is a case, and each case belongs to one of several ordinal categories. If there are only two categories, the ordinal model is identical to the Bernoulli model, where one category represents the cases and the other category represent the controls in the Bernoulli model. The cases in the ordinal model may be a sample from a larger population or they may constitute a complete set of observations. Ordinal data can be analyzed with the purely temporal, the purely spatial or the space-time scan statistics.

*Example:* For the ordinal model, the data may consist of everyone diagnosed with breast cancer during a ten-year period, with three different categories representing early, medium and late stage cancer at the time of diagnosis.

The ordinal model requires information about the location of each case in each category. Separate locations may be specified for each case, or the data may be aggregated for states, provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc, with multiple cases in the same or different categories at each data location. To do a temporal or space-time analysis, it is necessary to have a time for each case as well.

With the ordinal model it is possible to search for high clusters, with an excess of cases in the high-valued categories, for low clusters with an excess of cases in the low-valued categories, or simultaneously for both types of clusters. Reversing the order of the categories has the same effect as changing the analysis from high to low and vice versa.

**Related Topics:** *Analysis Tab, Case File, Coordinates File, Likelihood Ratio Test, Methodological Papers, Probability Model Comparison.*

---

## Exponential Model

The exponential model<sup>8</sup> is designed for survival time data, although it could be used for other continuous type data as well. Each observation is a case, and each case has one continuous variable attribute as well as a 0/1 censoring designation. For survival data, the continuous variable is the time between diagnosis and death or depending on the application, between two other types of events. If some of the data is censored, due to loss of follow-up, the continuous variable is then instead the time between diagnosis and time of censoring. The 0/1 censoring variable is used to distinguish between censored and non-censored observations.

*Example:* For the exponential model, the data may consist of everyone diagnosed with prostate cancer during a ten-year period, with information about either the length of time from diagnosis until death or from diagnosis until a time of censoring after which survival is unknown.

When using the temporal or space-time exponential model for survival times, it is important to realize that there are two very different time variables involved. The first is the time the case was diagnosed, and that is the time that the temporal and space-time scanning window is scanning over. The second is the survival time, that is, time between diagnosis and death or for censored data the time between diagnosis and censoring. This is an attribute of each case, and there is no scanning done over this variable. Rather, we are interested in whether the scanning window includes exceptionally many cases with a small or large value of this attribute.

It is important to note, that while the exponential model uses a likelihood function based on the exponential distribution, the true survival time distribution must not be exponential and the statistical inference (p-value) is valid for other survival time distributions as well. The reason for this is that the randomization is not done by generating observations from the exponential distribution, but rather, by permuting the space-time locations and the survival time/censoring attributes of the observations.

**Related Topics:** *Likelihood Ratio Test, Analysis Tab, Probability Model Comparison, Methodological Papers.*

---

## Normal Model

The normal model<sup>10</sup> is designed for continuous data. For each individual or for each observation, called a case, there is a single continuous attribute that may be either negative or positive. The model can also be used for ordinal data when there are many categories. That is, different cases are allowed to have the same attribute value.

*Example:* For the normal model, the data may consist of the birth weight and residential census tract for all newborns, with an interest in finding clusters with lower birth weight. One individual is then a ‘case’. Alternatively, the data may consist of the average birth weight in each census tract. It is then the census tract that is the ‘case’, and it is important to use the weighted normal model, since each average will have a different variance due to a different number of births in each tract.

It is important to note that while the normal model uses a likelihood function based on the normal distribution, the true distribution of the continuous attribute must not be normal. The statistical inference (p-value) is valid for any continuous distribution. The reason for this is that the randomization is not done by generating simulated data from the normal distribution, but rather, by permuting the space-time locations and the continuous attribute (e.g. birth weight) of the observations. While still being formally valid, the results can be greatly influenced by extreme outliers, so it may be wise to truncate such observations before doing the analysis.

In the standard normal model<sup>9</sup>, it is assumed that each observation is measured with the same variance. That may not always be the case. For example, if an observation is based on a larger sample in one location and a smaller sample in another, then the variance of the uncertainty in the estimates will be larger for the smaller sample. If the reliability of the estimates differs, one should instead use the weighted normal scan statistic<sup>10</sup> that takes these unequal variances into account. The weighted version is obtained in SaTScan by simply specifying a weight for each observation as an extra column in the input file. This weight may for example be proportional to the sample size used for each estimate or it may be the inverse of the variance of the observation.

If all values are multiplied with or added to the same constant, the statistical inference will not change, meaning that the same clusters with the same log likelihoods and p-values will be found. Only the estimated means and variances will differ. If the weight is the same for all observations, then the weighted normal scan statistic will produce the same results as the standard normal version. If all the weights are multiplied by the same constant, the results will not change.

**Related Topics:** *Analysis Tab, Likelihood Ratio Test, Methodological Papers, Probability Model Comparison.*



---

## Continuous Poisson Model

All the models described above are based on data observed at discrete locations that are considered to be non-random, as defined by a regular or irregular lattice of location points. That is, the locations of the observations are considered to be fixed, and we evaluate the spatial randomness of the observation conditioning on the lattice. Hence, those are all versions of what are called discrete scan statistics<sup>174</sup>. In a continuous scan statistics, observations may be located anywhere within a study area, such as a square or rectangle. The stochastic aspect of the data consists of these random spatial locations, and we are interested to see if there are any clusters that are unlikely to occur if the observations were independently and randomly distributed across the study area. Under the null hypothesis, the observations follow a homogeneous spatial Poisson process with constant intensity throughout the study area, with no observations falling outside the study area.

*Example:* The data may consist of the location of bird nests in a square kilometer area of a forest. The interest may be to see whether the bird nests are randomly distributed spatially, or in other words, whether there are clusters of bird nests or whether they are located independently of each other.

In SaTScan, the study area can be any collection of convex polygons, which are convex regions bounded by any number straight lines. Triangles, squares, rectangles, rhombuses, pentagons and hexagons are all examples of convex polygons. In the simplest case, there is only one polygon, but the study area can also be the union of multiple convex polygons. If the study area is not convex, divide it into multiple convex polygons and define each one separately. The study area does not need to be contiguous, and may for example consist of five different islands.

The analysis is conditioned on the total number of observations in the data set. Hence, the scan statistic simply evaluates the spatial distribution of the observation, but not the number of observations.

The likelihood function used as the test statistic is the same as for the Poisson model for the discrete scan statistic, where the expected number of cases is equal to the total number of observed observations, times the size of the scanning window, divided by the size of the total study area. As such, it is a special case of the variable window size scan statistic described by Kulldorff (1997)<sup>1</sup>. When the scanning window extends outside the study area, the expected count is still based on the full size of the circle, ignoring the fact that some parts of the circle have zero expected counts. This is to avoid strange non-circular clusters at the border of the study area. Since the analysis is based on Monte Carlo randomizations, the p-values are automatically adjusted for these boundary effects. The reported expected counts are based on the full circle though, so the Obs/Exp ratios provided should be viewed as a lower bound on the true value whenever the circle extends outside the spatial study region.

The continuous Poisson model can only be used for purely spatial data. It uses a circular scanning window of continuously varying radius up to a maximum specified by the user. Only circles centered on one of the observations are considered, as specified in the coordinates file. If the optional grid file is provided, the circles are instead centered on the coordinates specified in that file. The continuous Poisson model has not been implemented to be used with an elliptic window.

**Related Topics:** *Analysis Tab, Likelihood Ratio Test, Methodological Papers, Poisson Model, Probability Model Comparison.*

---

## Probability Model Comparison

In SaTScan, there are seven different probability models for discrete scan statistics. For count data, there are three different probability models: discrete Poisson, Bernoulli and space-time permutation. The ordinal and multinomial models are designed for categorical data with and without an inherent ordering from for example low to high. There are two models for continuous data: normal and exponential. The latter is primarily designed for survival type data. For continuous scan statistics there is only the homogeneous Poisson model.

The discrete Poisson model is usually the fastest to run. The ordinal model is typically the slowest.

With the discrete Poisson and space-time permutations models, an unlimited number of covariates can be adjusted for, by including them in the case and population files. With the normal model, it is also possible to adjust for covariates by including them in the case file, but only for purely spatial analyses. With the Bernoulli, ordinal, exponential and normal models, covariates can be adjusted for by using multiple data sets, which limits the number of covariate categories that can be defined, or through a pre-processing regression analysis done before running SaTScan.

All discrete probability models can be used for either individual locations or aggregated data.

With the discrete Poisson model, population data is only needed at selected time points and the numbers are interpolated in between. A population time must be specified even for purely spatial analyses. Regardless of model used, the time of a case or control need only be specified for purely temporal and space-time analyses.

The space-time permutation model automatically adjusts for purely spatial and purely temporal clusters. For the discrete Poisson model, purely temporal and purely spatial clusters can be adjusted for in a number of different ways. For the Bernoulli, ordinal, exponential and normal models, spatial and temporal adjustments can be done using multiple data sets, but it is limited by the number of different data sets allowed, and it is also much more computer intensive.

Purely temporal and space-time analyses cannot be performed using the homogeneous Poisson model. Spatial variation in temporal trend analyses can only be performed using the discrete Poisson model.

### ***Few Cases Compared to Controls***

In a purely spatial analysis where there are few cases compared to controls, say less than 10 percent, the discrete Poisson model is a very good approximation to the Bernoulli model. The former can then be used also for 0/1 Bernoulli type data, and may be preferable as it has more options for various types of adjustments, including the ability to adjust for covariates specified in the case and population files. As an approximation for Bernoulli type data, the discrete Poisson model produces slightly conservative p-values.

### ***Bernoulli versus Ordinal Model***

The Bernoulli model is mathematically a special case of the ordinal model, when there are only two categories. The Bernoulli model runs faster, making it the preferred model to use when there are only two categories.

### **Normal versus Exponential Model**

Both the normal and exponential models are meant for continuous data. The exponential model is primarily designed for survival time data but can be used for any data where all observations are positive. It is especially suitable for data with a heavy right tail. The normal model can be used for continuous data that takes both positive and negative values. While still formally valid, results from the normal model are sensitive to extreme outliers.

### **Normal versus Ordinal Model**

The normal model can be used for categorical data when there are very many categories. As such, it is sometimes a computationally faster alternative to the ordinal model. There is an important difference though. With the ordinal model, only the order of the observed values matters. For example, the results are the same for ordered values '1 – 2 – 3 – 4' and '1 – 10 – 100 – 1000'. With the normal model, the results will be different, as they depend on the relative distance between the values used to define the categories.

### **Discrete versus Homogeneous Poisson Model**

Instead of using the homogeneous Poisson model, the data can be approximated by the discrete Poisson model by dividing the study area into many small pieces. For each piece, a single coordinates point is specified, the size of the piece is used to define the population at that location and the number of observations within that small piece of area is the number of cases in that location. As the number of pieces increases towards infinity, and hence, as their size decreases towards zero, the discrete Poisson model will be asymptotically equivalent to the homogeneous Poisson model.

### **Temporal Data**

For temporal and space-time data, there is an additional difference among the probability models, in the way that the temporal data is handled. With the Poisson model, population data may be specified at one or several time points, such as census years. The population is then assumed to exist between such time points as well, estimated through linear interpolation between census years. With the Bernoulli, space-time permutation, ordinal, exponential and normal models, a time needs to be specified for each case and for the Bernoulli model, for each control as well.

**Related Topics:** *Bernoulli Model, Poisson Model, Space-Time Permutation Model, Likelihood Ratio Test, Methodological Papers.*

---

## **Likelihood Ratio Test**

For each location and size of the scanning window, the alternative hypothesis is that there is an elevated risk within the window as compared to outside. Under the Poisson assumption, the likelihood function for a specific window is proportional to<sup>1</sup>:

$$\left(\frac{c}{E[c]}\right)^c \left(\frac{C-c}{C-E[c]}\right)^{C-c} IO$$

where C is the total number of cases, c is the observed number of cases within the window and E[c] is the covariate adjusted expected number of cases within the window under the null-hypothesis. Note that

since the analysis is conditioned on the total number of cases observed,  $C-E[c]$  is the expected number of cases outside the window.  $I()$  is an indicator function. When SaTScan is set to scan only for clusters with high rates,  $I()$  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, then  $I()=1$  for all windows.

The space-time permutation model uses the same function as the Poisson model. Due to the conditioning on the marginals, the observed number of cases is only approximately Poisson distributed. Hence, it is no longer a formal likelihood ratio test, but it serves the same purpose as the test statistic.

For the Bernoulli model the likelihood function is<sup>1,2</sup>:

$$\left(\frac{c}{n}\right)^c \left(\frac{n-c}{n}\right)^{n-c} \left(\frac{C-c}{N-n}\right)^{C-c} \left(\frac{(N-n)-(C-c)}{N-n}\right)^{(N-n)-(C-c)} I()$$

where  $c$  and  $C$  are defined as above,  $n$  is the total number of cases and controls within the window, while  $N$  is the combined total number of cases and controls in the data set.

The likelihood function for the multinomial, ordinal, exponential, and normal models are more complex, due to the more complex nature of the data. We refer to papers by Jung, Kulldorff and Richards<sup>6</sup>, Jung, Kulldorff and Klassen<sup>7</sup>; Huang, Kulldorff and Gregorio<sup>8</sup>; Kulldorff et al<sup>9</sup>, and Huang et al.<sup>10</sup> for the likelihood functions for these models. The likelihood function for the spatial variation in temporal trends scan statistic is also more complex, as it involves the maximum likelihood estimation of several different trend functions.

The likelihood function is maximized over all window locations and sizes, and the one with the maximum likelihood constitutes the most likely cluster. This is the cluster that is least likely to have occurred by chance. The likelihood ratio for this window constitutes the maximum likelihood ratio test statistic. Its distribution under the null-hypothesis is obtained by repeating the same analytic exercise on a large number of random replications of the data set generated under the null hypothesis. The p-value is obtained through Monte Carlo hypothesis testing<sup>14</sup>, by comparing the rank of the maximum likelihood from the real data set with the maximum likelihoods from the random data sets. If this rank is  $R$ , then  $p = R / (1 + \text{\#simulation})$ . In order for  $p$  to be a ‘nice looking’ number, the number of simulations is restricted to 999 or some other number ending in 999 such as 1999, 9999 or 99999. That way it is always clear whether to reject or not reject the null hypothesis for typical cut-off values such as 0.05, 0.01 and 0.001.

The SaTScan program scans for areas with high rates (clusters), for areas with low rates, or simultaneously for areas with either high or low rates. The latter should be used rather than running two separate tests for high and low rates respectively, in order to make correct statistical inference. The most common analysis is to scan for areas with high rates, that is, for clusters.

### ***Non-Compactness Penalty Function***

When the elliptic window shape is used, there is an option to use a non-compactness (eccentricity) penalty to favor more compact clusters<sup>12</sup>. The main reason for this is that the elliptic scan statistic will under the null hypothesis typically generate an elliptic most likely cluster since there are more elliptic than circular clusters evaluated, and it will often be a long and narrow ellipse, since there are more of those. At the same time, the concept of clustering is based on a compactness criterion in the sense that the cases in the cluster should be close to each other, so we are more interested in compact clusters. When the non-compactness penalty is used, the pure likelihood ratio is no longer used as the test statistic.

Rather, the test statistic is defined as the log likelihood ratio multiplied with a non-compactness penalty of the form  $[4s/(s+1)^2]^a$ , where  $s$  is the elliptic window shape defined as the ratio of the length of the longest to the shortest axis of the ellipse. For the circle,  $s=1$ . The parameter  $a$  is a penalty tuning parameter. With  $a=0$ , the penalty function is always 1 irrespective of  $s$ , so that there is never a penalty. When  $a$  goes to infinity, the penalty function goes to 0 for all  $s>1$ , so that only circular clusters are considered. Other than this, there is no clear intuitive meaning of the penalty tuning parameter  $a$ . In SaTScan, it is possible to use either a strong penalty ( $a=1$ ) or a medium size penalty ( $a=1/2$ ).

**Related Topics:** *Batch Mode, Bernoulli Model, Covariate Adjustments, Elliptic Scanning Window, Exponential Model, Monte Carlo Replications, Ordinal Model, Poisson Model, Secondary Clusters, Space-Time Permutation Model, Standard Results File.*

---

## Secondary Clusters

For purely spatial and space-time analyses, SaTScan also identifies secondary clusters in the data set in addition to the most likely cluster, and orders them according to their likelihood ratio test statistic. There will almost always be a secondary cluster that is almost identical with the most likely cluster and that have almost as high likelihood value, since expanding or reducing the cluster size only marginally will not change the likelihood very much. Most clusters of this type 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. The user can decide to what extent overlapping clusters are reported in the results files. The default is that geographically overlapping clusters are not reported.

There may also be secondary clusters that do not overlap spatially with the most likely cluster, and they may be of great interest. These are always reported. The p-values for such clusters should be interpreted in terms of the ability of the secondary cluster to reject the null hypothesis on its own strength, whether or not the more likely clusters are true clusters or not. Hence, these p-values are not adjusted for the fact that there may be other clusters in the data. If such adjustments are desired, the iterative scan statistic should be used.

For purely temporal analyses, only the most likely cluster is reported.

**Related Topics:** *Adjusting for More Likely Clusters, Likelihood Ratio Test, Spatial Output Tab, Criteria for Reporting Secondary Clusters, Standard Results File.*

---

## Adjusting for More Likely Clusters

When there are multiple clusters in the data set, the secondary clusters are evaluated as if there were no other clusters in the data set. That is, they are statistically significant if and only if they are able to cause a rejection of the null hypothesis on their own strength, whether or not the other clusters are true clusters or not. That is often the desired type of inference. Sometime though, it is also of interest to evaluate secondary clusters after adjusting for other clusters in the data.

As an advanced option, SaTScan is able to adjust the inference of secondary clusters for more likely clusters in the data<sup>24</sup>. This is done in an iterative manner. In the first iteration SaTScan runs the standard analysis but only reports the most likely cluster. That cluster is then removed from the data set, including

all cases and controls (Bernoulli model) in the cluster while the population (Poisson model) is set to zero for the locations and the time period defining the cluster. In a second iteration, a completely new analysis is conducted using the remaining data. This procedure is then repeated until there are no more clusters with a p-value less than a user specified maxima or until a user specified maximum number of iterations have been completed, whichever comes first.

For purely spatial analyses it has been shown that the resulting p-values for secondary clusters are quite accurate and at most marginally biased.

Note that the circle of a secondary cluster may overlap with the circle of a previously detected more likely cluster, and it may even completely encircle it so that the latter is a subset of the former. This does not mean that the more likely cluster is detected twice. Rather, the more likely cluster is treated as a 'lake' with no population and no cases, and the new secondary cluster consist of the areas around that 'lake'. This may for example happen if a city has a very high elevated risk, while the surrounding suburbs have a modest elevated risk. The same phenomena may occur when doing purely temporal or space-time analyses.

This feature is not available for the continuous Poisson model.

**Related Topics:** *Spatial Output Tab, Criteria for Reporting Secondary Clusters, Iterative Scan, Likelihood Ratio Test, Secondary Clusters, Standard Results File.*

---

## Covariate Adjustments

A covariate should be adjusted for when all three of the following are true:

- The covariate is related to the disease in question.
- The covariate is not randomly distributed geographically.
- You want to find clusters that cannot be explained by that covariate.

Here are three examples:

- If you are studying cancer mortality in the United States, you should adjust for age since (i) older people are more likely to die from cancer (ii) some areas such as Florida have a higher percent older people, and (iii) you are presumably interested in finding areas where the risk of cancer is high as opposed to areas with an older population.
- If you are interested in the geographical distribution of birth defects, you do not need to adjust for gender. While birth defects are not equally likely in boys and girls, the geographical distribution of the two genders is geographically random at time of birth.
- If you are studying the geography of lung cancer incidence, you should adjust for smoking if you are interested in finding clusters due to non-smoking related risk factors, but you should not adjust for smoking if you are interested in finding clusters reflecting areas with especially urgent needs to launch an anti-smoking campaign.

When the disease rate varies, for example, with age, and the age distribution varies in different areas, then there is geographical clustering of the disease simply due to the age covariate. When adjusting for categorical covariates, the SaTScan program will search for clusters above and beyond that which is

expected due to these covariates. When more than one covariate is specified, each one is adjusted for as well as all the interaction terms between them.

**Related Topics:** *Covariate Adjustment Using the Input Files, Covariate Adjustment using Statistical Regression Software, Covariate Adjustment Using Multiple Data Sets, Methodological Papers.*

## Covariate Adjustment Using the Input Files

With the Poisson and space-time permutation models, it is possible to adjust for multiple categorical covariates by specifying the covariates in the input files. To do so, simply enter the covariates as extra columns in the case file (both models) and the population file (Poisson model). There is no need to enter any information on any of the window tabs.

For the Poisson model, the expected number of cases in each area under the null-hypothesis is calculated using indirect standardization. Without covariate adjustment the expected number of cases in a location is (spatial analysis):

$$E[c] = p * C / P$$

where  $c$  is the observed number of cases and  $p$  the population in the location of interest, while  $C$  and  $P$  are the total number of cases and population respectively. Let  $c_i$ ,  $p_i$ ,  $C_i$  and  $P_i$  be defined in the same way, but for covariate category  $i$ . The indirectly standardized covariate adjusted expected number of cases (spatial analysis) is:

$$E[c] = \sum_i E[c_i] = \sum_i p_i * C_i / P_i$$

The same principle is used when calculating the covariate adjusted number of cases for the space-time scan statistic, although the formula is more complex due to the added time dimension.

Since the space-time permutation model automatically adjusts for purely spatial and purely temporal variation, there is no need to adjust for covariates in order to account for different spatial or temporal densities of these covariates. For example, there is no need to adjust for age simply because some places have a higher proportion of old people. Rather, covariate adjustment is used if there is space-time interaction due to this covariate rather than to the underlying disease process. For example, if children get sick mostly in the summer and adults mostly in the winter, then there will be age generated space-time interaction clusters in areas with many children in the summer and vice versa. When including child/adult as a covariate, these clusters are adjusted away.

**Note:** Too many covariate categories can create problems. For the space-time permutation model, the adjustment is made at the randomization stage, so that each covariate category is randomized independently. If there are too many covariate categories, so that all or most cases in a category belong to the same spatial location or the same aggregated time interval, then there is very little to randomize, and the test becomes meaningless.

**Related Topics:** *Covariate Adjustments, Covariate Adjustment using Statistical Regression Software, Covariate Adjustment Using Multiple Data Sets, Methodological Papers, Poisson Model, Space-Time Permutation Model, Case File, Population File.*

## Covariate Adjustment Using Statistical Regression Software

SaTScan cannot in itself do an adjustment for continuous covariates. Such adjustments can still be done for the Poisson model<sup>17,23</sup>, but it is a little more complex. The first step is to calculate the covariate adjusted expected number of cases for each location ID and time using a standard statistical regression software package like SAS. These expected numbers should then replace the raw population numbers in the population file, while not including the covariates themselves.

The use of external regression software is also an excellent way to adjust for covariates in the exponential model<sup>8</sup>. The first step is to fit an exponential regression model without any spatial information, in order to obtain risk estimates for each of the covariates. The second step is to adjust the survival and censoring time up or down for each individual based on the risk estimates his or her covariates.

For the normal model, covariates can be adjusted for by first doing linear regression using standard statistical software, and then replacing the observed value with their residuals.

**Related Topics:** *Covariate Adjustments, Covariate Adjustment Using the Input Files, Covariate Adjustment Using Multiple Data Sets, Exponential Model, Methodological Papers, Poisson Model, Population File.*

## Covariate Adjustment Using Multiple Data Sets

It is also possible to adjust for categorical covariates using multiple data sets<sup>11</sup>. The cases and controls/population are then divided into categories, and a separate data set is used for each category. This type of covariate adjustment is computationally much slower than the one using the input files, and is not recommended for large data sets. One advantage is that it can be used to adjust for covariates when running the multinomial or ordinal models, for which other adjustment procedures are unavailable. A disadvantage is that since the maximum number of data sets allowed by SaTScan is twelve, the maximum number of covariate categories is also twelve.

The adjustment approach to multiple data sets is as follows (when searching for clusters with high rates):

1. For each window location and size, the log likelihood ratio is calculated for each data set.
2. The log likelihood ratio for all data sets with less than expected number of cases in the window is multiplied with negative one.
3. The log likelihood ratios are then summed up, and this sum is the combined log likelihood for that particular window.
4. The maximum of all the combined log likelihood ratios, taken over all the window locations and sizes, constitutes the most likely cluster, and this is evaluated in the same way as for a single data set.

When searching for clusters with low rates, the same procedure is performed, except that it is then the data sets with more than expected cases that we multiply by one. When searching for both high and low clusters, both sums are calculated, and the maximum of the two is used to represent the log likelihood ratio for that window.



**Related Topics:** *Multiple Data Sets Tab, Covariate Adjustment, Covariate Adjustment Using the Input Files, Covariate Adjustment using Statistical Regression Software, Methodological Papers, Bernoulli Model.*

---

## Spatial and Temporal Adjustments

### Adjusting for Temporal Trends

If there is an increasing temporal trend in the data, then the temporal and space-time scan statistics will pick up that trend by assigning a cluster during the end of the study period. If there is a decreasing trend, it will instead pick up a cluster at the beginning of the time period. Sometimes it is of interest to test whether there are temporal and/or space-time clusters after adjusting for a temporal trend.

For the space-time permutation model, the analysis is automatically adjusted for both temporal trends and temporal clusters, and no further adjustments are needed. For the discrete Poisson model, the user can specify whether a temporal adjustment should be made, and if so, whether to adjust with a percent change or non-parametrically.

Sometimes, the best way to adjust for a temporal trend is by specifying the percent yearly increase or decrease in the rate that is to be adjusted for. This is a log linear adjustment. Depending on the application, one may adjust either for a trend that SaTScan estimates from the data being analyzed, or from the trend as estimated from national or other similar data. In the latter case, the percent increase or decrease must be calculated using standard statistical regression software such as SAS or R, and then inserted on the Risk Adjustments Tab.

For space-time analyses, it is also possible to adjust for a temporal trend non-parametrically. This adjusts the expected count separately for each aggregated time interval, removing all purely temporal clusters. The randomization is then stratified by time interval to ensure that each time interval has the same number of events in the real and random data sets.

The ability to adjust for temporal trends is much more limited for the Bernoulli, multinomial, ordinal, normal and exponential models, as none of the above features can be used. Instead, the time must be divided into discrete time periods, with the cases and controls in each period corresponding to a separate data set with separate case and control files. The analysis is then done using multiple data sets.

**Related Topics:** *Spatial and Temporal Adjustments Tab, Time Aggregation, Adjusting for Day-of-Week Effects, Poisson Model.*

### Adjusting for Day-of-Week Effects

Some data sets have a weekly pattern. If not adjusted for, that could create clusters, for example on a Monday, or from one Monday to the next Monday, simply because Mondays in general have more events than other days of the week. One way to adjust for this is to aggregate daily data into weeks, but that will reduce the temporal resolution. Another option is to select the day-of-week adjustment feature on the Spatial and Temporal Adjustment Tab, which will non-parametrically adjust for any weekly adjustment in the data. This feature is only available with the discrete Poisson probability model.

The space-time permutation model automatically adjusts for any purely temporal variation in the data, including day-of-week effects. Hence, with this probability model, there is never a need to do any special day-of-week adjustment. If different spatial locations have different day-of-week effects, that may lead to spurious space-time interaction clusters. For example, if disease data comes from different medical clinics, but only some of the clinics are open on the weekends. That may result in weekend clusters at those clinics that are simply an artifact of their opening hours. To adjust for this, it is possible to adjust for the space-by-day-of-week interaction by selecting that option on the Spatial and Temporal Adjustment Tab. Doing this has exactly the same effect as including a day-of-week variable in the input case file. This feature is only valid for the space-time permutation model.

**Related Topics:** *Spatial and Temporal Adjustments Tab, Adjusting for Temporal Trends, Time Aggregation, Poisson Model.*

## Adjusting for Purely Spatial Clusters

In a space-time analysis with the Poisson model, it is also possible to adjust for purely spatial clusters, in a non-parametric fashion. This adjusts the expected count separately for each location, removing all purely spatial clusters. The randomization is then stratified by location ID to ensure that each location has the same number of events in the real and random data sets.

This option is not available for the Bernoulli, multinomial, ordinal, exponential, normal or space-time permutation models, in the latter case because the method automatically adjusts for any purely spatial clusters.

**Note:** It is not possible to simultaneously adjust for spatial clusters and purely temporal clusters using stratified randomization, and if both types of adjustments are desired, the space-time permutation model should be used instead.

**Related Topics:** *Spatial and Temporal Adjustments Tab, Poisson Model, Adjusting for Temporal Trends.*

## Adjusting for Known Relative Risks

Sometimes it is known a priori that a particular location and/or time has a higher or lower risk of known magnitude, and we want to detect clusters above and beyond this, or in other words, we want to adjust for this known excess/lower risk. One way to do this is to simply change the population at risk numbers in the population file. A simpler way is to use the adjustments file. In this file, a relative risk is specified for any location and time period combination. The expected counts are then multiplied by this relative risk for that location and time. For example, if it is known from historical data that a particular location typically have 50 percent more cases during the summer months June to August, then for each year one would specify a relative risk of 1.5 for this location and these months. A summer cluster will then only appear in this location if the excess risk is more than 50 percent.

This feature is only available for the discrete Poisson model.

**Related Topics:** *Adjustments File, Spatial and Temporal Adjustments Tab, Time Aggregation, Poisson Model, Missing Data*

---

## Missing Data

If there is missing data for some locations and times, it is important to adjust for that in the analysis. If not, you may find statistically significant low rate clusters where there is missing data, or statistically significant high rate clusters in other locations, even though these are simply artifacts of the missing data.

### ***Bernoulli Model***

To adjust a Bernoulli model analysis for missing data, do the following. If cases are missing for a particular location and time period remove the controls for that same location and time. Likewise, if controls are missing for a particular location and time, remove the cases for that same location and time. This needs to be done before providing the data to SaTScan. If both cases and controls are missing for a location and time, you are fine, and there is no need for any modification of the input data.

### ***Multinomial and Ordinal Models***

To adjust a multinomial or ordinal model analysis for missing data, do the following. If one or more categories are missing for a particular location and time period, remove all cases in the remaining categories from that same location and time. This needs to be done before providing the data to SaTScan. If all cases in all categories are missing for a location and time, you are fine, and there is no need for any modification of the input data.

### ***Discrete Poisson Model***

To adjust the discrete Poisson model for missing data, use the adjustments file to define the location and time combinations for which the data is missing, and assign a relative risk of zero to those location/time combinations.

### ***Continuous Poisson Model***

To adjust the continuous Poisson model for missing data, redefine the study area by using a different set of polygons, so that areas with missing data are excluded from the study area.

### ***Space-Time Permutation Model***

It is a little more complex to adjust for missing data in the space-time permutation model, but still possible<sup>5</sup>. First add day-of-week as a covariate in the analysis file. When a particular location / time period is missing, then for that location, remove all data for the days of the week for which any data is missing. For example, if data from Thursday 10/23 and Friday 10/24 are missing for zip-code area A and data from Saturday 10/25 are missing from area B, remove data from all Thursdays and Fridays for area A and data from all Saturdays from area B, while retaining all data from Saturdays through Wednesdays for area A and all data except Saturdays from area B. For all other zip code areas, retain all data for all days. Note that, in addition to adjusting for the missing data, this approach will also adjust for any day-of-week by spatial interaction effects.

The same approach can be used with other categorization of the data, as long as the categorizations is in some time-periodic unit that occur several times and is evenly spread out over the study period. For example, it is okay to categorize into months if the study period spans several years, but not if you only have one year's worth of data.

Two more crude approaches to deal with missing data in the space-time permutation model is to remove all data for a particular location if some data are missing for that location or to remove all data for a particular time period for dates on which there is missing data in any location. The latter is especially useful in prospective surveillance for missing data during the beginning of the study period, to avoid removing recent data that are the most important for the early detection of disease outbreaks.

**Note:** When there are location/time combinations with missing data, either remove the whole row from the case file or assign zero cases to that location/time combination. If you only remove the number of cases, but retain the location ID and time information, there will be a file reading error.

**Warning:** The adjustment for missing data only works if the locations and times for which the data is missing is independent of the number of cases in that location and time. For example, if data is missing for all locations with less than five observed cases, the adjustment procedures described above will not work properly.

**Related Topics:** *Adjustments File, Adjusting for Known Relative Risks, Bernoulli Model, Ordinal Model, Poisson Model, Space-Time Permutation Model, Spatial and Temporal Adjustments Tab, Time Aggregation*

---

## Multivariate Scan with Multiple Data Sets

Sometimes it is interesting to simultaneously search for and evaluate clusters in more than one data set. For example, one may be interested in spatial clusters with excess incidence of leukemia only, of lymphoma only or of both simultaneously. As another example, one may be interested in detecting a gastrointestinal disease outbreak that affects children only, adults only or both simultaneously. If SaTScan is used to analyze one single combined data set, one may miss a cluster that is only present in one of the subgroups. On the other hand, if two SaTScan analyses are performed, one for each data set, there is a loss of power if the true cluster is about equally strong in both data sets. A SaTScan analysis with multiple data sets and the multivariate scan option solves this problem.

The multivariate scan statistic with multiple data sets works as follows<sup>12</sup> (when searching for clusters with high rates):

1. For each window location and size, the log likelihood ratio is calculated for each data set.
2. The log likelihood ratios for the data sets with more than expected number of cases is summed up, and this sum is the likelihood for that particular window.
3. The maximum of all the summed log likelihood ratios, taken over all the window locations and sizes, constitutes the most likely cluster, and this is evaluated in the same way as for a single data set.

When searching for clusters with low rates, the same procedure is performed, except that we instead sum up the log likelihood ratios of the data sets with fewer than expected number of cases within the window in question. When searching for both high and low clusters, both sums are calculated, and the maximum of the two is used to represent the log likelihood ratio for that window.

**Note:** All data sets must use the same probability model and the same geographical coordinates file.

**Related Topics:** *Multiple Data Sets Tab, Covariate Adjustment Using Multiple Data Sets, Coordinates File.*

# Comparison with Other Methods

---

## Scan Statistics

Scan statistics were first studied in detail by Joseph Naus<sup>184</sup>. A major challenge with scan statistics is to find analytical results concerning the probabilities of observing a cluster of a specific magnitude and there is a beautiful collection of mathematical theory that has been developed to obtain approximations and bounds for these probabilities under a variety of settings. Excellent reviews of this work have been provided by Glaz and Balakrishnan<sup>174</sup>, Glaz, Naus and Wallenstein<sup>175</sup> and Glaz, Pozdnyakov and Wallenstein<sup>246</sup>. Two common features of this early work on scan statistics were: (i) they use a fixed size scanning window, and (ii) they deal with count data where under the null hypothesis, the observed number of cases follow a uniform distribution in either a continuous or discrete setting, so that the expected number of cases in an area is proportional to the size of that area.

In disease surveillance, neither of these assumptions is met, since we do not know the size of a cluster a priori and since the population at risk is geographically inhomogeneous. Under the null hypothesis of equal disease risk one expects to see more disease cases in a city compared to a similar sized area in the countryside, just because of the higher population density in the city. The scan statistics in the SaTScan software were developed to resolve these two problems. Since no analytical solutions have been found to obtain the probabilities under these more complex settings, Monte Carlo hypothesis testing is instead used to obtain the p-values<sup>14</sup>.

---

## Spatial and Space-Time Clustering

### *Descriptive Cluster Detection Methods*

In 1987, Openshaw et al.<sup>185</sup> developed a Geographical Analysis Machine (GAM) that uses overlapping circles of different sizes in the same way as the spatial scan statistic, except that the circle size does not vary continuously. With the GAM, a separate significance test is made for each circle, leading to multiple testing, and in almost any data set there will be a multitude of ‘significant clusters’ when defined in this way. This is because under the null hypothesis, each circle has a 0.05 probability of being ‘significant’ at the 0.05 level, and with 20,000 circles we would expect 1,000 ‘significant’ clusters under the null-hypothesis of no clusters. GAM is hence very useful for descriptive purposes, but should not be used for hypothesis testing.

Another nice method for descriptive cluster detection was proposed by Rushton and Lolonis<sup>187</sup>, who used p-value contour maps to depict the clusters rather than overlapping circles. As with GAM, it does not adjust for the multiple testing inherent in the many potential cluster locations evaluated.

### *Cluster Detection Tests*

The spatial scan statistic is a cluster detection test. A cluster detection test is able to both detect the location of clusters and evaluate their statistical significance without problems with multiple testing. In 1990, Turnbull et al.<sup>191</sup> proposed the first such test using overlapping circles with fixed population size, assigning the circle with the most cases as the detected cluster.

The spatial scan statistic was in part inspired by the work of Openshaw et al.<sup>185</sup> and Turnbull et al.<sup>191</sup>. By applying a likelihood ratio test, it was possible to evaluate clusters of different sizes (as Openshaw et al. did) while at the same time adjusting for the multiple testing (as Turnbull et al. did).

In a power comparison<sup>2</sup>, it was shown that Turnbull's method has higher power if the true cluster size is within about 20 percent of what is specified by that method, while the spatial scan statistic has higher power otherwise. Note that the cluster size in Turnbull's method must be specified before looking at the data, or the procedure is invalid.

### ***Focused Cluster Tests***

Focused tests should be used when there is a priori knowledge about the location of the hypothesized cluster. For example, a cluster around a toxic waste site in one country may spur an investigation about clusters around a similar toxic waste site in another country. The spatial scan statistic or other cluster detection tests should then not be used, as they will have low power due to the evaluation of all possible locations even though the hypothesized location is already known. Examples of focused tests are Stone's Test<sup>188</sup>, Lawson-Waller's Score Test<sup>181, 192</sup> and Bithell's Test<sup>170</sup>.

Focused tests should never be used when the foci were defined using the data itself. This would lead to pre-selection bias and the resulting p-values would be incorrect. It is then better to use the spatial scan statistic. If on the other hand, the point source was defined without looking at the data, then it is better to use the focused test rather than the spatial scan statistic, as the former will have higher power as it focuses on the location of interest.

In addition to various scan statistics, the SaTScan software can also be used to do a focused test in order to evaluate whether there is a disease cluster around a pre-determined focus (ref. 2, p809). This is done by using a grid file with only a single grid point reflecting the coordinates of the focus of interest. Similarly, a multi-focused test can be specified using the grid file with one coordinate for each desired focus.

### ***Global Clustering Tests***

Most proposed tests for spatial clustering are tests for global clustering. These include among many others the methods proposed by Alt and Vach<sup>167</sup>, Besag and Newell<sup>169</sup>, Cuzick and Edwards<sup>171</sup>, Diggle and Chetwynd<sup>172</sup>, Grimson<sup>176</sup>, Moran<sup>183</sup>, Ranta<sup>186</sup>, Tango<sup>189, 190</sup>, Walter<sup>193</sup> and Whittemore et al.<sup>194</sup>. These methods test for clustering throughout the study region without the ability to pinpoint the location of specific clusters. As such, these tests and the spatial scan statistic complement each other, since they are useful for different purposes.

### ***Global Space-Time Interaction Tests***

Knox<sup>178</sup>, Mantel<sup>182</sup>, Diggle et al.<sup>173</sup>, Jacquez<sup>177</sup>, Baker<sup>168</sup>, and Kulldorff and Hjalmars<sup>180</sup>, have proposed different tests for space-time interaction. Like the space-time permutation<sup>5</sup> version of the space-time scan statistic, these methods are designed to evaluate whether cases that are close in space are also close in time and vice-versa, adjusting for any purely spatial or purely temporal clustering. Being global in nature, these other tests are useful when testing to see if there is clustering throughout the study region and time period, and the preferred method when for example trying to determine whether a disease is infectious. Unlike the space-time permutation based scan statistic though, they are unable to detect the location and size of clusters and to test the significance of those clusters.

**Related Topics:** *Likelihood Ratio Test, SaTScan Methodology Papers*

# Input Data

---

## Data Requirements

**Required Files:** The input data should be provided in a number of files. A coordinates file is always needed and a case file is needed for all probability models except the continuous Poisson model. The Poisson model also requires a population file while the Bernoulli model requires a control file.

**Optional Files:** One may also specify an optional special grid file that contains geographical coordinates of the centroids defining the circles used by the scan statistic. If such a file is not specified, the coordinates in the coordinate file will be used for that purpose. As part of the advanced features, there is also an optional max circle size file, an optional adjustments file, and optional non-Euclidian neighbors file and an optional meta location file.

**File Format:** The data input files must be in SaTScan ASCII file format or you may use the SaTScan import wizard for dBase, comma delimited or space delimited files. Using such files, the wizard will automatically generate SaTScan file format files. Both options are described below.

**Spatial Resolution:** For the discrete scan statistics, separate data locations may be specified for individuals or data may be aggregated for states, provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc.

**Temporal Information:** To do a temporal, a space-time or a spatial variation in temporal trends analysis, it is necessary to have a time related to each case, and if the Bernoulli model is used, for each control as well. This time can be specified as a day, month or year. When the discrete Poisson model is used the background denominator population is assumed to exist continuously over time, although not necessarily at a constant level. The population file requires a date to be specified for each population count. For times in-between those dates, SaTScan will estimate the population through linear interpolation. If all population counts have the same date, the population is assumed to be constant over time.

**Multiple Data Sets:** It is possible to specify multiple case files, each representing a different data set, with information about different diseases or about men versus women respectively. For the Bernoulli model, each case file must be accompanied with its own control file, and for the Poisson model, each case file must be accompanied with its own population file. The maximum number of data sets that SaTScan can analyze is twelve.

**Covariate Adjustments:** With the Poisson and space-time permutation models, it is possible to adjust for multiple categorical covariates by including them in the case and population files. For the Bernoulli, ordinal or exponential models, covariates can be adjusted for using multiple data sets.

**Related Topics:** *Input Tab, Multiple Data Sets Tab, Case File, Control File, Population File, Coordinates File, Grid File, SaTScan Import Wizard, SaTScan ASCII File Format, Covariate Adjustments.*



---

## Case File

The case file provides information about cases, and it is used for all probability models. It should contain the following information:

**Location ID:** Any numerical value or string of characters. Empty spaces may not form part of the id.

**Number of Cases:** The number of cases for the specified location, time and covariates. For the discrete Poisson, binomial and space-time permutation models, this is the number of observations or individuals with the characteristic of interest, such as cancer or low birth weight. For the ordinal, multinomial, normal and exponential models, it is the total number of observations or individuals in the locations, irrespectively of the value of their categorical characteristic or continuous attribute value.

**Date/Time:** Optional. May be specified either in years, months or days, or in a generic format. The format must coincide with the time precision format specified on the Input Tab. Unless temporal data check is disabled, all case times must fall within the study period as specified on the Input Tab.

**Attribute:** For the multinomial, ordinal, exponential and normal models only. A variable describing some characteristic of the case. These may be a category (multinomial or ordinal model), survival time (exponential model), or a continuous variable value (normal model). The categories for the multinomial and ordinal models can be specified as any positive or negative numerical value. Survival times must be positive numbers. The numbers for the normal model can be positive or negative.

**Censored:** For the exponential model only. Censored is a 0/1 variable with censored=1 and uncensored=0.

**Weight:** Optional. For the normal model only. Required if covariates are used, even if all observations have the same variance, in which case all weights should be set to one.

**Covariates:** Optional. For discrete Poisson, space-time permutation and normal models only. Any number of categorical covariates may be specified as either numbers or through characters. For the normal model, covariates can only be included if weights are also provided.

**Example:** If on April 1, 2004 there were 17 male and 12 female cases in New York, the following information would be provided:

NewYork 12 2004/4/1 Female

NewYork 17 2004/4/1 Male

**Note:** For the weighted normal model, there can be only one case (observation) per line, and hence, if weights are specified, the second column must be all ones.

**Note:** Multiple lines may be used for different cases with the same location, time and attributes. SaTScan will automatically add them.

**Note:** This file is not used for the continuous Poisson model.

**Related Topics:** *Input Tab, Case File Name, Multiple Data Sets Tab, Covariate Adjustment Using Input Files, SaTScan Import Wizard, SaTScan ASCII File Format.*

---

## Control File

The control file is only used with the Bernoulli model. It should contain the following information:

**location id:** Any numerical value or string of characters. Empty spaces may not form part of the id.

**#controls:** The number of controls for the specified location and time.

**time:** Optional. Time may be specified either in years, months or days, or in a generic format. All control times must fall within the study period as specified on the Analysis tab. The format of the times must be the same as in the case file.

**Note:** Multiple lines may be used for different controls with the same location, time and attributes. SaTScan will automatically add them.

**Related Topics:** *Input Tab, Control File Name, Multiple Data Sets Tab, SaTScan Import Wizard, SaTScan ASCII File Format.*

---

## Population File

The population file is used for the discrete Poisson model, providing information about the background population at risk. This may be actual population count from a census, or it could be for example covariate adjusted expected counts from a statistical regression model. It should contain the following information:

**location id:** Any numerical value or string of characters. Empty spaces may not form part of the id.

**time:** The time to which the population size refers. May be specified either in years, months or days, or in a generic format. If the population time is unknown but identical for all population numbers, then a dummy year must be given, the choice not affecting result.

**population:** Population size for a particular location, year and covariate combination. If the population size is zero for a particular location, year, and set of covariates, then it should be included in the population file specified as zero. The population can be specified as a decimal number to reflect a population size at risk rather than an actual number of people.

**covariates:** Optional. Any number of categorical covariates may be specified, each represented by a different column separated by empty spaces. May be specified numerically or through characters. The covariates must be the same as in the case file.

**Example:** If age and sex are the covariates included, with 18 different age groups, then there should be  $18 \times 2 = 36$  rows for each year and census area. With 3 different census years, and 32 census areas, the file will have a total of 3456 rows and 5 columns.

**Note:** Multiple lines may be used for different population groups with the same location, time and covariate attributes. SaTScan will automatically add them.

**Note:** For a purely temporal analysis with the discrete Poisson model, it is not necessary to specify a population file if the population is constant over time.

**Related Topics:** *Input Tab, Population File Name, Multiple Data Sets Tab, Covariate Adjustment Using Input Files, Max Circle Size File, SaTScan Import Wizard, SaTScan ASCII File Format.*

---

## Coordinates File

The coordinates file provides the geographic coordinates for each location ID. Each line of the file represents one geographical location. Area-based information may be aggregated and represented by one single geographical point location. Coordinates may be specified either using the standard Cartesian coordinate system or in latitude and longitude. If two different location IDs have exactly the same coordinates, then the data for the two are combined and treated as a single location.

A coordinates file is not needed for purely temporal analyses.

**Related Topics:** *Input Tab, Coordinates File Name, Coordinates, Cartesian Coordinates, Latitude and Longitude, Grid File.*

## Cartesian Coordinates

Cartesian is the mathematical name for the regular planar x,y-coordinate system taught in high school. These may be specified in two, three or any number of dimensions. The SaTScan program will automatically read the number of dimensions, which must be the same for all coordinates. If Cartesian coordinates are used, the coordinates file should contain the following information:

**location id:** Any numerical value or string of characters. Empty spaces may not form part of the id.

**coordinates:** The coordinates must all be specified in the same units. There is no upper limit on the number of dimensions.

**x and y-coordinates:** Required

**z1-zN coordinates:** Optional

**Note:** If you have more than 10 dimensions you cannot use the SaTScan Import Wizard for the coordinates and grid files, but must specify them using the SaTScan ASCII file format.

**Note:** The continuous Poisson model only works in two dimensions.

**Related Topics:** *Input Tab, Coordinates, Latitude and Longitude, Coordinates File, Grid File, SaTScan Import Wizard, SaTScan ASCII File Format.*

## Latitude and Longitude

Latitudes and longitudes should be entered as decimal number of degrees. Latitude represents the north/south distance from the equator, and locations south of the equator should be entered as negative numbers. Longitude represents the east/west distance from the Prime Meridian (Greenwich, England), and locations west of the Prime Meridian should be entered as negative numbers. For example, the

National Institutes of Health in Bethesda, Maryland, which is located at 39.00 degrees north and 77.10 degrees west, should be reported as 39.00 and -77.10 respectively.

Latitudes and longitudes can, for the purpose of this program, not be specified in degrees, minutes and seconds. Such latitudes and longitudes can easily be converted into decimal numbers of degrees (DND) by the simple formula:  $DND = \text{degrees} + \text{minutes}/60 + \text{seconds}/3600$ .

If latitude/longitude coordinates are used, the coordinates file should contain the following information:

**location id:** Any numerical value or string of characters. Empty spaces may not form part of the id.

**latitude:** Latitude in decimal number of degrees.

**longitude:** Longitude in decimal number of degrees.

**Note:** When coordinates are specified in latitudes and longitudes, SaTScan does not perform a projection of these coordinates onto a planar space. Rather, SaTScan draws perfect circles on the surface of the spherical earth.

**Note:** Latitude and longitude cannot be used for the continuous Poisson model, or when an elliptic spatial window is used.

**Related Topics:** *Input Tab, Coordinates File, Coordinates, Cartesian Coordinates, Latitude and Longitude, Grid File, SaTScan Import Wizard, SaTScan ASCII File Format, Computing Time.*

---

## Grid File

The optional grid file defines the centroids of the circles used by the scan statistic. If no grid file is specified, the coordinates given in the coordinates file are used for this purpose. Each line in the file represents one circle centroid. There should be at least two variables representing Cartesian (standard) x,y-coordinates or exactly two variables representing latitude and longitude. The choice between Cartesian and latitude/longitude must coincide with the coordinates file, as must the number of dimensions.

The grid file will normally only include spatial coordinates, while the temporal range of potential clusters is specified on the Temporal Window Tab. That does not allow the user to specify a different temporal range for different locations. If that is desired, four more columns can be added to the grid file representing the earliest allowed start time of the cluster, the latest allowed start time, the earliest allowed end time and the latest allowed end time, in that order. If these columns are specified for some grid points, but not for remainder, then the remaining grid points will have the temporal cluster specifications defined on the Temporal Window Tab.

If only one centroid is specified in the grid file, one gets a focused cluster test rather than a scan statistic. Such focused tests are useful to evaluate whether there is a cluster around a pre-defined location such as a toxic waste site. If more than one but still a small number of centroids is specified in the grid file, one gets a multi-focused tests, looking for clusters around one or more of the centroids.

**Related Topics:** *Input Tab, Grid File Name, Coordinates, Cartesian Coordinates, Latitude and Longitude, Coordinates File, SaTScan Import Wizard, SaTScan ASCII File Format, Temporal Window Tab, Computing Time.*

---

## Non-Euclidian Neighbors File

This is an optional file for the discrete scan statistics. It cannot be defined using the SaTScan Import Wizard, but has to be specified using the ASCII file format. With this option, the coordinates and grid files are not needed, and ignored if provided.

With the standard parameter settings, SaTScan uses the coordinates file to determine which locations are closest to the center of each circle constructed. This is done using Euclidean distances. In essence, for each centroid SaTScan finds the closest neighbor, the second closest, and so on, until it reaches the maximum window size. With the neighbors file, it is possible for the user to specify these neighbor relations in any way without being constrained to Euclidean distances. For example, the neighbors may be sorted according to distance along a subway network or a water distribution system.

The first column of this file contains the location IDs defining the centroids of the scanning window. The subsequent entries on each row are then the centroids neighbors in order of closeness. The scanning window will expand in size until there are no more neighbors provided for that row. That means that this file also defines the maximum window size. It is allowed to have multiple rows for the same location ID centroid, each with a different set of closest neighbors.

**Note:** The neighbors file cannot be used with the continuous Poisson model.

**Related Topics:** *Coordinates File, Input Tab, Meta Location File, Spatial Neighbors Tab, SaTScan ASCII File Format.*

---

## Meta Location File

This is an optional file for the discrete scan statistics which can only be used if the non-Euclidian neighbors file is used as well. It cannot be defined using the SaTScan Import Wizard, but has to be specified using the ASCII file format.

A meta location is a collection of two or more individual location IDs. When a meta location is specified in the non-Euclidian neighbors file, all individual members of the meta location is simultaneously entered into the scanning window.

The first column of this file contains the user defined names of the meta locations. The subsequent entries on each row are the individual location IDs that are part of that meta location. There is no upper limit on the number of individual locations that can belong to each meta location.

**Note:** The meta location file can only be used in connection with the non-Euclidian neighbors file.

**Related Topics:** *Coordinates File, Input Tab, Spatial Neighbors Tab, Non-Euclidian Neighbors File, SaTScan ASCII File Format.*

---

## Max Circle Size File

This optional file is used to determine the maximum circle size of the scanning window, when the maximum is defined as a percentage of the ‘population’. Normally, the percentage is based on the population in the population file, but by using the max circle size file, a different ‘population’ can be specified for this purpose. One important reason for using the max circle size file is for prospective space-time analyses, where the regular population file may change over time, but one wants to evaluate the same set of geographical circles each time. This is critical in order to properly adjust the prospective space-time scan statistic for earlier analyses. It can also be used for other purposes.

The file should contain one line for each location, with the following information:

**location id:** Any numerical value or string of characters. Empty spaces may not form part of the id.

**‘population’:** Any non-negative number.

The name of the special max circle size file is specified on the Analysis Tab → Advanced Features → Spatial Window Tab.

**Note:** If a location ID is missing from this file, the population is assumed to be zero. If a location ID occurs more than once, the population numbers will be added.

**Related Topics:** *Input Tab, Population File, Spatial Window Tab, SaTScan Import Wizard, SaTScan ASCII File Format.*

---

## Adjustments File

The adjustments file can be used to adjust a discrete Poisson model analysis for any temporal, spatial and space-time anomalies in the data, with a known relative risk. It can for example be used to adjust for missing or partially missing data. (Note: Covariates are adjusted for by using the case and population files or by analyzing multiple data sets, not with this file). The adjustments file should contain one or more lines for each location for which adjustments are warranted, with the following information:

**Location ID:** Any numerical value or string of characters. Empty spaces may not form part of the id. Alternatively, it is possible to specify ‘All’, in which all location will be adjusted with the same relative risk.

**Relative Risk:** Any non-negative number. The relative risk representing how much more common disease is in this location and time period compared to the baseline. Setting a value of one is equivalent of not doing any adjustments. A value of greater than one is used to adjust for an increased risk and a value of less than one to adjust for lower risk. A relative risk of zero is used to adjust for missing data for that particular time and location.

**Start Time:** Optional. The start of the time period to be adjusted using this relative risk.

**End Time:** Optional. The end of the time period to be adjusted using this relative risk.

If no start and end times are given, the whole study period will be adjusted for that location. If 'All' is selected instead of a location ID, but no start or end times are given, that has the same effect as when no adjustments are done.

The name of the adjustments file is specified on the Analysis Tab → Advanced Features → Risk Adjustments.

**Note:** Assigning a relative risk of  $x$  to half the locations is equivalent to assigning a relative risk of  $1/x$  to the other half. Assigning the same relative risk to all locations and time periods has the same effect as not adjusting at all.

**Note:** It is permissible to adjust the same location and time periods multiple times, through different rows with different relative risks. SaTScan will simply multiply the relative risks. For example, if you adjust location A with a relative risk of 2 for all time periods, and you adjust 1990 with a relative risk of 2 for all locations, then the 1990 entry for location A will be adjusted with a relative risk of  $2*2=4$ .

**Related Topics:** *Adjustments with Known Relative Risk, Missing Data, Spatial and Temporal Adjustments Tab, SaTScan Import Wizard, SaTScan ASCII File Format.*

---

## Alternative Hypothesis File

When estimating statistical power, the alternative hypothesis file is used to define the alternative hypothesis for which the power is estimated. It has the same format as the adjustments file.

**Location ID:** Any numerical value or string of characters that is present in the geographical coordinates file. Empty spaces may not form part of the id. Alternatively, it is possible to specify 'All', in which all location will have the same relative risk under the alternative hypothesis.

**Relative Risk:** Any non-negative number. The relative risk representing how much more common disease is in this location and time period compared to the baseline. Setting a value of one is the same as not including that row at all. A value of greater than one is used to for alternative hypotheses with increased risk, and a value of less than one for alternative hypothesis with lower risk. It is fine to define an alternative hypothesis with different relative risks for different location IDs, and it is even okay to have an alternative hypothesis with a relative risk greater than 1 for some locations and less than 1 for other locations.

**Start Time:** Optional. The start of the time period that has a different relative risk in the alternative hypothesis.


**End Time:** Optional. The end of the time period that has a different relative risk in the alternative hypothesis.

It is possible to evaluate multiple alternative hypotheses within the same analysis run. This is done by leaving a blank row between the different alternative hypotheses. This is computationally more efficient than running a separate analysis for each alternative.

**Related Topics:** *Adjustments File, Power Estimation Tab.*

---

## SaTScan Import Wizard

The SaTScan Import Wizard can be used to import dBase, comma-delimited, or space-delimited files. It works for all import files except the optional Neighbors File. Launch the Import Wizard by clicking on the File Import  button to the right of the text field for the file that you want to import. Use the **Next** and **Previous** buttons to navigate between the dialogs. Follow the steps below to import files.

### ***Step 1 – Selecting the Source File***

1. At the bottom of the Select Source File dialog, select the file type extension you are looking for. If you are unsure, select the All Files option. Supported file formats are: dBase III/IV, CSV and Excel 97-2003
2. Browse the folders and highlight the file you want to open. It will appear in the File Name text field.
3. Click on Open. The SaTScan Import Wizard will now appear.

### ***Step 2: Specifying the File Format***

If you are importing a dBase or AN Excel file, this step is automatically skipped. For all other source files, you need to specify the file structure using the File Format dialog box.

1. First specify whether you have a character delimited or fixed column file format, using the radio buttons under the **Source File Type** heading.
2. If there are extraneous lines in the beginning of the file, type the number lines that you would like to ignore in the text field in the upper right corner.
3. If you have a character delimited file, use the scrolling menus to select the field separator to be either a comma, a semicolon or white space.
4. If you have a fixed column file, define the fields using the Field Information box. For each field type the name, the start column, and the length (maximum number of characters) into the appropriate spaces. Click on the **Add** button to add another field. The information will appear in the panel on the right. Continue adding fields until you have the appropriate number. To change the information in the right panel, highlight the line you want to change. The information will appear in the **Field Information** box. Edit the information and click on the **Update** button when you are done. The updated information will appear in the right panel.
5. Click on **Next** to proceed to the next dialog box.



### Step 3: Matching Source File Variables with SaTScan Variables

The top grid in this dialog box links the SaTScan variables with the input file variables from the source file. The bottom grid displays sample data from the chosen input file.

1. If there are headers in your file, click the checkbox in the lower left corner.
2. To match the variables, click on one of the places where it says 'unassigned'.
3. Select the appropriate variable from the input file to go with the chosen SaTScan variable.
4. When all the required and optional variables that you selected have been matched, click on the Execute button to import the file. This will create a temporary file in SaTScan ASCII file format.
5. If the input file has headings that are exactly the same as the SaTScan variable names, you can click on the Auto Align button to match these automatically.

When importing the case file, the variables to match varies depending on the probability model used. By selecting the probability model at the top of the import wizard will only display the variables relevant to that model.

### Step 4: Saving the Imported File

The imported file, which is in SaTScan ASCII file format, must be saved at least temporarily. The default is to save it to the TEMP directory and after the analysis is completed you may erase the file. You can also save it to some other directory of your choice and use it for future analyses without having to recreate it by using the Import Wizard again.

**Related Topics:** *Input Tab, Case File, Control File, Population File, Coordinates File, Grid File, Max Circle Size File, Adjustments File.*

---

## SaTScan ASCII File Format

As an alternative to using the SaTScan Import Wizard, it is also possible to directly write the name of the input files in the text fields provided on the Input Tab, or to browse the file directories for the desired input files using the button to the right of that box. The files must then be in SaTScan file format, which are space delimited ASCII files with one row for each location/covariate combination and with columns as defined below. Such files can be created using any text editor and most spreadsheets. The order of the columns in the file is very important, but the rows can be in any order. The optional variables, defined above, are optional columns in the SaTScan file format.

Case File Format (\*.cas):

**<location id> <#cases> <time> <attribute><censored><weight><covariate#1> ... <covariate#N>**

The use of attributes, censored, weight and covariates depends on the probability model, as shown in Table 1.

Probability Model	attribute	censored	weight	covariates
Discrete Poisson	n/a	n/a	n/a	optional
Bernoulli	n/a	n/a	n/a	n/a
Space-Time Permutation	n/a	n/a	n/a	optional
Multinomial	category	n/a	n/a	n/a
Ordinal	category	n/a	n/a	n/a
Exponential	survival time	optional	n/a	n/a
Normal	continuous variable	n/a	optional	optional

Table 1: Case file attributes used by the different probability models.

Control File Format (\*.ctl):

**<location ID> <#controls> <time>**

Population File Format (\*.pop):

**<location ID> <time> <population> <covariate#1> ... <covariate#N>**

Coordinates File Formats (\*.geo):

**<location ID> <latitude> <longitude> OR**

**<location ID> <x-coordinate> <y-coordinate> <z1-coordinate> ... <zN-coordinate>**

Grid File Formats (\*.grd):

**<latitude> <longitude> OR**

**<x-coordinate> <y-coordinate> <z1-coordinate> ... <zN-coordinate> OR**

**<latitude> <longitude><earliest start time><latest start time><earliest end time><latest end time>  
OR**

**<x-coordinate> <y-coordinate> <z1-coordinate> ... <zN-coordinate><earliest start time>**

**<latest start time><earliest end time><latest end time>**

Non-Euclidian Neighbors File Format (\*.nbr):

**<location ID> <location ID of closest neighbor> <location ID of 2<sup>nd</sup> closest neighbor> etc**

Meta Location File Format (\*.met):

**<meta location ID> <location ID #1> <location ID #2> etc**

Special Max Circle Size File Format (\*.max):

**<location ID> <'population'>**

Adjustment File Format (\*.adj):

**<location ID> <relative risk> <start time> <end time>**

Alternative Hypothesis File Format (\*.ha):

**<location ID> <relative risk> <start time> <end time>**

## ***Time Formats***

Times must be entered in a specific format. Generic time is specified using any negative or positive integer in the range (-200,000 to 2,900,000). If you have times outside this interval, simply add or subtract the same constant to all the times. The valid date formats are:

2010  
2010/06, 2010/06/26  
2010-06, 2010-06-26  
06/2010, 06/26/2010  
06-2010, 06-26-2010

Single digit days and months may be specified with one or two digits. For example, September 9, 2002, can be written as 2002/9/9, 2002/09/09, 2002/09/9, 2002/9/09, 2002-9-9, etc.

**Note:** SaTScan also support a few other time formats used in earlier versions, but they are no longer recommended.

**Related Topics:** *Input Tab, Case File, Control File, Population File, Coordinates File, Grid File, Max Circle Size File, Neighbors File, Adjustments File, SaTScan Import Wizard.*

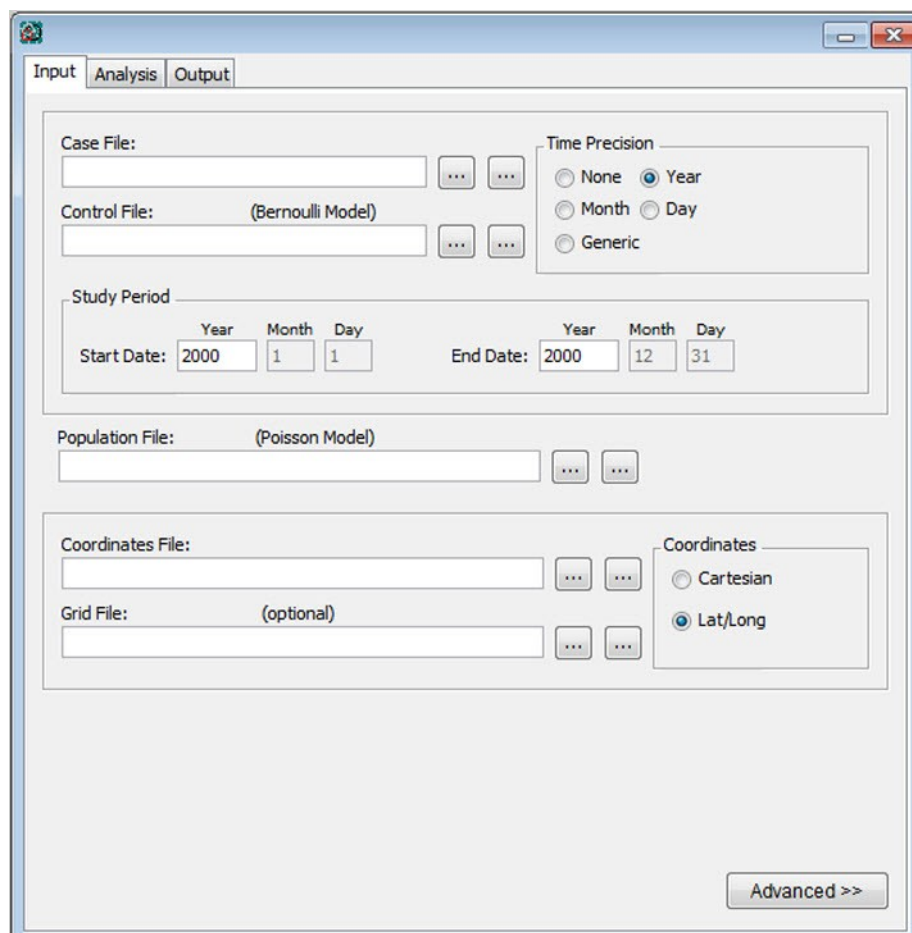
# Basic SaTScan Features

Most SaTScan analyses can be performed using the basic analysis and data features. The users specify these on three different window tabs for input, analysis and output options respectively. These contain all required specifications for a SaTScan analysis as well as a few optional ones. Additional features, all optional, can be specified on the advanced features tabs.

**Related Topics:** *Statistical Methodology, Input Tab, Analysis Tab, Output Tab, Advanced Features.*

---

## Input Tab



The screenshot shows the 'Input' tab of the SaTScan dialog box. It features three tabs: 'Input', 'Analysis', and 'Output'. The 'Input' tab is active. The dialog is organized into several sections: 'Case File' with a text box and a browse button; 'Control File' with a text box, a browse button, and a dropdown menu set to '(Bernoulli Model)'; 'Study Period' with 'Start Date' and 'End Date' each having 'Year', 'Month', and 'Day' input boxes (Start Date is 2000, 1, 1; End Date is 2000, 12, 31); 'Population File' with a text box, a browse button, and a dropdown menu set to '(Poisson Model)'; 'Coordinates File' with a text box and a browse button; 'Grid File' with a text box, a browse button, and a dropdown menu set to '(optional)'; and 'Coordinates' with radio buttons for 'Cartesian' and 'Lat/Long' (selected). A button labeled 'Advanced >>' is located at the bottom right.

***Input Tab Dialog Box***

The Input Tab is used to specify the names of the input data files as well as the nature of the data in these files. If the files are in SaTScan ASCII file format, they may be specified either by writing the name in the text box or by using the browse button [...]. If they are not in SaTScan ASCII file format, they must be specified using the SaTScan import wizard, by clicking on the File Import [...] button. Both the SaTScan ASCII file format and the SaTScan import wizard are described in the Input Data section.

**Related Topics:** *Basic SaTScan Features, Input Data, Multiple Data Sets Tab.*

## Case File Name

Specify the name of the input file with case data. This file is required for all discrete scan statistics, irrespective of the probability model used.

**Related Topics:** *Input Tab, Case File.*

## Control File Name

Specify the name of the input file with control data. This file is only used for analyses with the Bernoulli probability model.

**Related Topics:** *Input Tab, Control File.*

## Time Precision

Indicate whether the case file and the control file (when applicable) contain information about the time of each case (and control), and if so, whether the precision should be read as generic days, months or years. If the time precision is specified to be days but the precision in the case or control file is in month or year, then there will be an error. If the time precision is specified as years, but the case or control file includes some dates specified in terms of the month or day, then the month or day will be ignored.

For a purely spatial analysis, the case and control file need not contain any times. If they do, it has to be specified that they do contain this information so that SaTScan knows how to read the file, but the information is ignored.

**Note:** The choice defines only the precision for the times in the case and control files. The precision of the times in the population file can be different, except that if one has generic times the other must also have generic times.

**Related Topics:** *Input Tab, Case File, Control File, Study Period, Time Aggregation.*

## Study Period

Specify the start and end date of the time period under study. This must be done even for a purely spatial analysis in order to calculate the expected number of cases correctly. Allowable years are those between 1753 and 9999.

All times in the case and control files should fall on or between the start and end date of the study period. Dates in the population file are allowed to be outside the start and end date of the study period.

**Start Date/Time:** The earliest date/time to be included in the study period.

**End Date/Time:** The latest date/time to be included in the study period.

**Note:** The start and end dates cannot be specified to a higher precision than the precision of the times in the case and control files.

If the user does not specify month, then by default it will be set to January for the start date and to December for the end date. Likewise, if day is not specified, then by default it will be set to the first of the month for the start date and the last of the month for the end date.

**Related Topics:** *Input Tab, Case File, Control File, Time Precision, Time Aggregation.*

## **Population File Name**

Specify the name of the input file with population data. This file is only used for analyses using the discrete Poisson probability model.

**Related Topics:** *Input Tab, Population File.*

## **Coordinates File Name**

Specify the name of the input file with geographical coordinates of all the locations with data on the number of cases, controls and/or population. When multiple data sets are used, the coordinates file must include the coordinates for all locations found in any of the data sets.

**Related Topics:** *Input Tab, Coordinates, Coordinates File.*

## **Grid File Name**

Specify the name of the optional grid file with the coordinates of the circle centroids used by the spatial and space-time scan statistics. If no special grid file is specified, then the coordinates in the coordinates file are used for this purpose.

**Related Topics:** *Input Tab, Coordinates, Coordinates File, Grid File.*

## **Coordinates**

Specify the type of coordinates used by the coordinates file and the grid file, as either Cartesian or latitude/longitude. Cartesian is the mathematical name for the regular x/y-coordinate system taught in high school. Latitude/longitude cannot be used for the continuous Poisson model.

**Related Topics:** *Cartesian Coordinates, Latitude/Longitude, Coordinates File, Grid File.*

---

## Analysis Tab

The Analysis Tab dialog box contains the following settings:

- Type of Analysis:**
  - Retrospective Analyses:
    - ☒ Purely Spatial
    - ☐ Purely Temporal
    - ☐ Space-Time
    - ☐ Spatial Variation in Temporal Trends
  - Prospective Analyses:
    - ☐ Purely Temporal
    - ☐ Space-Time
- Probability Model:**
  - Discrete Scan Statistics:
    - ☒ Poisson
    - ☐ Bernoulli
    - ☐ Space-Time Permutation
    - ☐ Multinomial
    - ☐ Ordinal
    - ☐ Exponential
    - ☐ Normal
  - Continuous Scan Statistics:
    - ☐ Poisson ...
- Scan For Areas With:**
  - ☒ High Rates
  - ☐ Low Rates
  - ☐ High or Low Rates
- Time Aggregation:**
  - Units: ☒ Year, ☐ Month, ☐ Day
  - Length: 1 Years

Advanced >>

*Analysis Tab Dialog Box*

The Analysis Tab is used to set various analysis options. Additional features are available by clicking on the Advanced button in the lower right corner.

**Related Topics:** *Basic SaTScan Features, Statistical Methodology, Spatial Window Tab, Temporal Window Tab, Spatial and Temporal Adjustments Tab, Inference Tab.*

### Type of Analysis

SaTScan may be used for a purely spatial, purely temporal, space-time analyses and spatial variation in temporal trends. A purely spatial analysis ignores the time of cases, even when such data are provided. A purely temporal analysis ignores the geographical location of cases, even when such information is provided.

Purely temporal and space-time data can be analyzed in either retrospective or prospective fashion. In a retrospective analysis, the analysis is done only once for a fixed geographical region and a fixed study period. SaTScan scans over multiple start dates and end dates, evaluating both ‘alive clusters’, lasting until the study period and date, as well as ‘historic clusters’ that ceased to exist before the study period

end date. The prospective option is used for the early detection of disease outbreaks, when analyses are repeated every day, week, month or year. Only alive clusters, clusters that reach all the way to current time as defined by the study period end date, are then searched for.

**Related Topics:** *Spatial Temporal and Space-Time Scan Statistics, Analysis Tab, Methodological Papers, Computing Time, Spatial Window Tab, Temporal Window Tab, Time Aggregation.*

## Probability Model

There are eight different probability models that can be used: discrete Poisson, Bernoulli, space-time permutation, multinomial, ordinal, exponential, normal and continuous Poisson. For purely spatial analyses, the Poisson and Bernoulli models are good approximations for each other in many situations. Temporal data are handled differently, so the models differ more for temporal and space-time analyses.

**Discrete Poisson Model:** The discrete Poisson model should be used when the background population reflects a certain risk mass such as total person years lived in an area. The cases are then included as part of the population count.

**Bernoulli Model:** The Bernoulli model should be used when the data set contains individuals who may or may not have a disease and for other 0/1 type variables. Those who have the disease are cases and should be listed in the case file. Those without the disease are 'controls', listed in the control file. The controls could be a random set of controls from the population, or better, the total population except for the cases. The Bernoulli model is a special case of the ordinal model when there are only two categories.

**Space-Time Permutation Model:** The space-time permutation model should be used when only case data is available, and when one wants to adjust for purely spatial and purely temporal clusters.

**Multinomial Model:** The multinomial model is used when individuals belong to one of three or more categories, and when there is no ordinal relationship between those. When there are only two categories, the Bernoulli model should be used instead.

**Ordinal Model:** The ordinal model is used when individuals belong to one of three or more categories, and when there is an ordinal relationship between those categories such as small, medium and large. When there are only two categories, the Bernoulli model should be used instead.

**Exponential Model:** The exponential model is used for survival time data, to search for spatial and/or temporal clusters of exceptionally short or long survival. The survival time is a positive continuous variable. Censored survival times are allowed for some but not all individuals.

**Normal Model:** The normal model is used for continuous data. Observations may be either positive or negative.

**Continuous Poisson Model:** The continuous Poisson model should be used when the null hypothesis is that observations are distributed randomly with constant intensity according to a homogeneous Poisson process over a user defined study area.

**Related Topics:** *Analysis Tab, Bernoulli Model, Exponential Model, Methodological Papers, Ordinal Model, Poisson Model, Probability Model Comparison, Space-Time Permutation Model, .*



## Polygons for the Continuous Poisson Model

For the continuous Poisson model it is necessary to define the spatial study area in which the point observations may be located. This is done using one or more convex polygons, where each polygon is defined by a number of linear inequalities. For example, the unit square is defined by  $y \geq 0$ ,  $y \leq 1$ ,  $x \geq 0$  and  $x \leq 1$ . The study area is the unit, or sum, of all the areas defined by the different polygons. There is no upper limit on the number of polygons that can be used, nor on the number of inequalities used for each polygon. This means that almost any study area can be approximated to whatever precision wanted. The smallest number of inequalities that can be used to define a polygon is three, in which case the polygon is a triangle.

A new polygon is defined by first clicking the add button on the left to add a polygon, and then clicking the add button on the right to add a linear inequality. The first inequality is then specified using the equation editor at the bottom, followed by a click on the update button. After that, another inequality is added, and so on, until all the polygons have been defined. If you need to change an inequality, use the mouse to highlight the inequality you want to change, make the desired change in the equation editor, and then click on the update button.

**Note:** The polygons must be non-overlapping. They do not need to be contiguous.

**Related Topics:** *Analysis Tab, Continuous Poisson Model.*

## Scan for High or Low Rates

It is possible to scan for areas with high rates only (clusters), for areas with low rates only, or simultaneously for areas with either high or low rates. The most common analysis is to scan for areas with high rates only, that is, for clusters. For the exponential model, high corresponds to short survival. For the ordinal and normal models, high corresponds to large value categories/observations. By default, the multinomial model will simultaneously evaluate high and low rates for all categories. With the continuous Poisson model, it is only possible to scan for high rates.

The spatial variation in temporal trends scan statistic is not looking for clusters with either high or low rates. Rather, it is looking for ‘clusters’ with a trend that is higher or lower than the trend outside the cluster. As with the other scan statistics, it can look for clusters with high trends only, low trends only or simultaneously for both types. A cluster can have a high trend either because it has a rate that is increasing more than outside the cluster or because it has a rate that is decreasing less than outside the cluster. Likewise, a cluster can have a low trend either because it has a rate that is increasing less than outside the cluster or because it has a rate that is decreasing more than outside the cluster.

**Related Topics:** *Analysis Tab, Likelihood Ratio Test, Methodological Papers.*

## Time Aggregation

Space-time analyses are sometimes very computer intensive. To reduce the computing time, case times may be aggregated into time intervals. Another reason for doing so is to adjust for cyclic temporal trends. For example, when using intervals of one year, the analysis will automatically be adjusted for seasonal variability in the counts, and when using time intervals of 7 days, it will automatically adjust for weekday effects.

**Units:** The units in which the length of the time intervals are specified. This can be in years, months, days or generic. The units of the time intervals cannot be more precise than the time precision specified on the input tab. If generic time is used in the case file the unit for time aggregation must also be in generic time, as vice versa.

**Length:** The length of the time intervals in the specified units.

**Example:** If interval units are years and the length is two, then the time intervals will be two years long.

**Note:** If the time interval length is not a fraction of the length of the whole study period, the earliest time interval will be the remainder after the other intervals have received their proper length. Hence, the first time interval may be shorter than the specified length. For a spatial variation in temporal trends analysis, all time intervals must be of equal length, and if the first time interval is shorter, a warning is generated and that interval is ignored in the analysis.

**Important:** For prospective space-time analyses, the time interval must be equal to the length between the time-periodic analyses performed. So, if the time-period analyses are performed every week, then the time interval should be set to 7 days.

**Related Topics:** *Analysis Tab, Time Precision, Study Period, Computational Speed.*

---

## Output Tab

	ASCII	dBase
Cluster Information	<input type="checkbox"/>	<input type="checkbox"/>
Stratified Cluster Information	<input type="checkbox"/>	<input type="checkbox"/>
Location Information	<input type="checkbox"/>	<input type="checkbox"/>
Risk Estimates for Each Location	<input type="checkbox"/>	<input type="checkbox"/>
Simulated Log Likelihood Ratios/Test Statistics	<input type="checkbox"/>	<input type="checkbox"/>

***Output Tab Dialog Box***

Use the Output Tab is used to set parameters defining the output information provided by SaTScan.

**Related Topics:** *Results of Analysis, Standard Results File, Text Output Format, Column Output Format, Spatial Output Tab.*

### Text Output Format

A standard text based results file is automatically shown after the completion of the calculations. It contains information about the clusters detected, summary information about the data, computing time and the analysis parameters chosen. Specify the name of this file. Other optional output files may also be created, but must be opened manually by the user. These will have the same name, but with other filename extensions.

**Warning:** If you specify the name of a file that already exists, the old file will be overwritten and lost.

**Related Topics:** *Output Tab, Column Output Format, Geographical Output Format, Temporal Graph Output File.*

## Geographical Output Format

**KML File:** For spatial and space-time analyses, SaTScan will create a KML file that will show the detected clusters in Google Earth and other geographical software. In addition to the cluster location and size, the KML file also contains basic information about each cluster such as the observed number of cases, the relative risk and the p-value. If you select this option, SaTScan will automatically launch Google Earth and show you the results when the analysis is completed. If you do not want to have this automatic launch, you can deselect it on the Advanced Output Tab, and instead simply click on the KML file when you want to show the results. For this to work, you need to have the free Google Earth software installed on your computer. There are also other geographical software packages that can read KML files.

**Shape File:** For spatial and space-time analyses, SaTScan will create a Shape file that can be used to depict the detected clusters in geographical information systems. Two different files are created with extensions .shp and .shx.

The names of these output files are the same as the names of the text output format files, but with different filename extensions.

**Related Topics:** *Output Tab, Results of Analysis, Column Format Output Files, Cluster Information File, Location Information File, Temporal Graph Output File.*

## Column Output Format

In addition to the standard results file that is automatically shown at the completion of the calculations, it is possible to request five additional output files in column format with different types of information. These are useful for importing the results into other software.

- **Cluster Information File:** One row for each cluster, with information about that cluster.
- **Stratified Cluster Information File:** For each cluster, there is one row for each data set when multiple input data sets are used, and there is one row for each category used by the multinomial or ordinal model. For each cluster, data set and category, the file contains observed and expected cases, their ratio and the relative risk. This file is only useful for the multinomial and ordinal models or when there are multiple data sets. For other analyses this file is redundant as it contains a subset of the information already in the Cluster Information File.
- **Location Information File:** One row for each location ID, with information about that location and its cluster membership.
- **Risk Estimates File:** One row for each location ID, with the estimated risk in that location.
- **Simulated Log Likelihood Ratios File:** One row for each simulated data set, with the log likelihood ratio test statistic for that data set. This file is primarily used by statisticians interested in the distributional properties of scan statistics.

You must manually open all these files after the run is completed. They are provided in either ASCII or dBase format so that they can be easily imported into spreadsheets, geographical information systems or other database software. They have the same name as the standard text based output file, but with a different filename extension.

**Related Topics:** *Output Tab, Results of Analysis, Cluster Information File, Column Headers, Location Information File, Risk Estimates for Each Location, Geographical Output Format, Temporal Graph Output File.*

# Advanced Features

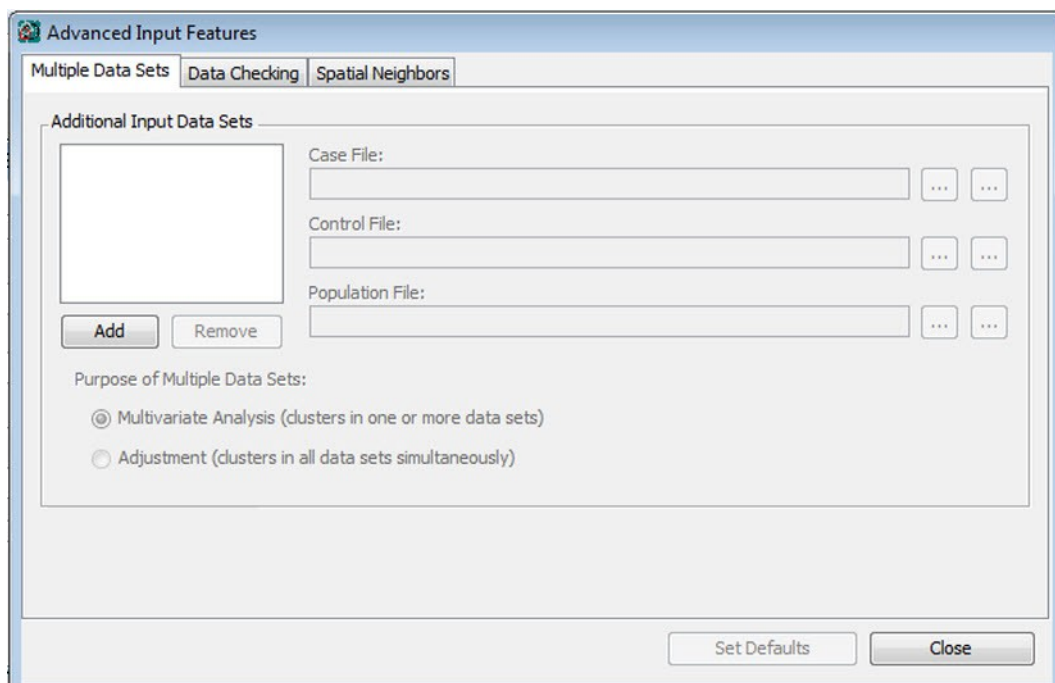
While most SaTScan analyses can be performed using the features on the three basic tabs for input, analysis and output parameters, additional options are warranted for some types of analyses, and these are available as advanced features. These features are reached through the Advanced button on the lower right corner of each of the three main tabs. ‘Advanced’ should be interpreted as ‘additional’ or ‘uncommon’ rather than ‘complex’, ‘difficult’ or ‘better’.

Since many of the advanced options depend on the selections made on the Input and Analysis Tabs, it is recommended that those two tabs be filled in first.

**Related Topics:** *Basic SaTScan Features, Multiple Data Sets Tab, Spatial Window Tab, Temporal Window Tab, Spatial and Temporal Adjustments Tab, Inference Tab, Spatial Output Tab.*



---

## Multiple Data Sets Tab



***Multiple Data Sets Tab Dialog Box***

It is possible to search and evaluate clusters in multiple data sets, as described in the Statistical Methodology section. The first data set is defined on the Input Tab. Up to eleven additional data sets can be defined on the Multiple Data Sets Tab. These files must be of the same class as the first one. That is, if the first data set consists of a case and a control file, so must all the others as well. The time precision and study period must also be the same as on the Input Tab.

Data sets are added by first clicking on the “Add” button, and then entering the file names by either typing it in the text box, by using the browser button  or through the SaTScan Import Wizard, Import File  button. Remove a data set by selecting it and clicking on the “Remove” button.

Multiple data sets can be used for two different purposes. One purpose is when there are different types of data, and we want to know if there is a cluster in either one or more of the data sets. The evidence for a cluster could then come exclusively from one data set or it may use the combined evidence from two or more data sets. The other purpose is to adjust for covariates. In this case the evidence of a cluster is based on all data sets. The difference is discussed in more detail in the statistical methodology section.

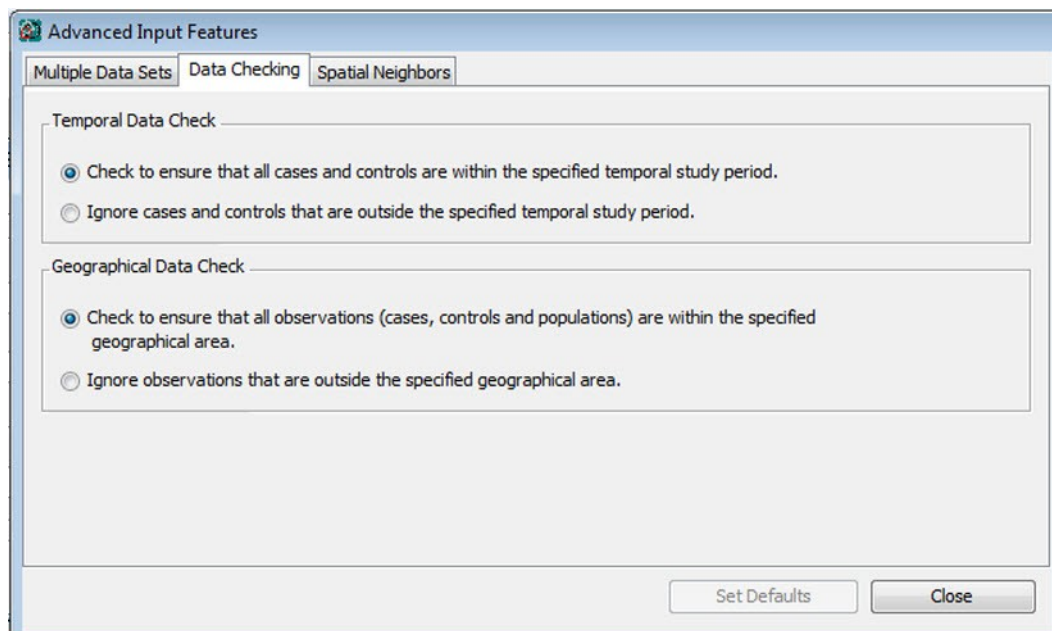
**Note:** Multiple data sets cannot be used for the continuous Poisson model.

**Warning:** The computing time is considerably longer when analyzing multiple data sets as compared to a single data set. Hence, it is not recommended to use multiple data sets when there are many locations in the coordinates file.

**Related Topics:** *Advanced Features, Input Tab, Multivariate Scan with Multiple Data Sets, Covariate Adjustments Using Multiple Data Sets, Computing Time, Case File, Control File, Population File.*

---

## Data Checking Tab



*Data Checking Tab Dialog Box*

## Temporal Data Check

By default, SaTScan will check that all the cases and all the controls are within the specified temporal study period. On this tab, it is possible to turn this off. Cases and controls outside the study period will then be ignored. This may be used if, for example, you only want to analyze a temporal subset of the data in the case and control input files.

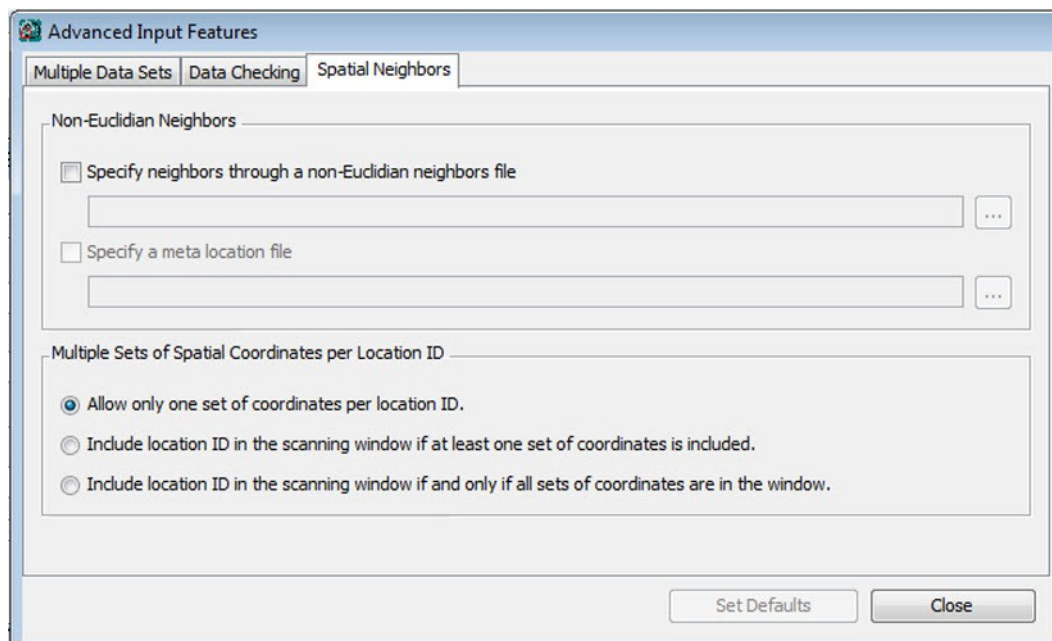
## Geographical Data Check

By default, SaTScan will check that all the cases, controls and population numbers are within the geographical area specified. For the discrete scan statistic, this means that they must be at one of the locations specified in the coordinates file. For the continuous scan statistics, it means that all the coordinates specified in the coordinates file must be within the polygons specified. On this tab, it is possible to turn off this data checking procedure. Data outside the geographical study area are then ignored. This may be used if, for example, you only want to analyze a geographical subset of the data, in which case only the geographical coordinates file has to be modified for a discrete scan statistic while the other files can be used as they are.

**Related Topics:** *Advanced Features, Case File, Input Tab, Study Time Period.*

---

## Spatial Neighbors Tab



*Spatial Neighbors Tab Dialog Box*



## Non-Euclidian Neighbors File

Rather than using circles or ellipses defined by the Euclidean distances between the locations specified in the coordinates and grid files, it is possible to manually specify a neighborhood matrix, to define neighbors by non-Euclidian distance. For each centroid, its closest, 2<sup>nd</sup> closest, 3<sup>rd</sup> closest neighbors are specified in turn and so on. This option is activated by checking the box on this tab and specifying the name of the neighbors file containing the neighbor matrix information. The format of the neighbors file is described in the ASCII File format section.

## Meta Location File

A meta location is a collection of two or more individual location IDs. When a meta location is specified in the special neighbors file, all individual members of the meta location is simultaneously entered into the scanning window. This option is activated by checking the box on this tab and specifying the name of the meta location file containing information about the individual location IDs belong to each meta location. The format of the neighbors file is described in the ASCII File format section.

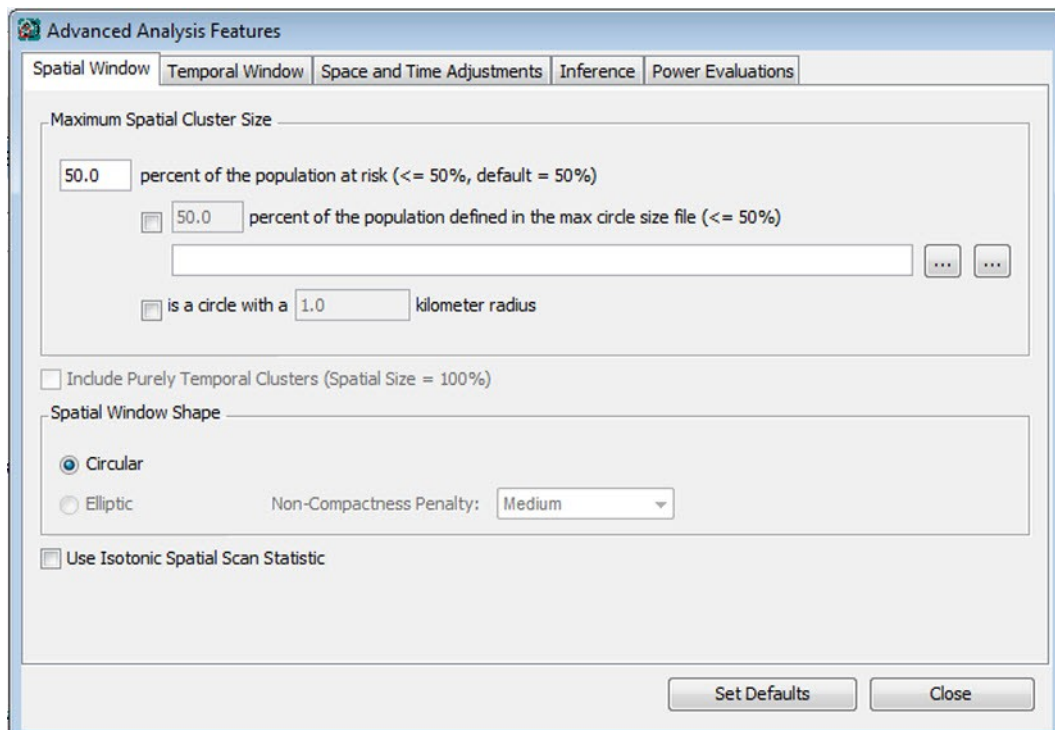
## Multiple Coordinates per Location

Each location ID is normally defined by a single set of coordinates, such as an (x,y) pair or a latitude and longitude. As an advanced option, it is possible to define multiple sets of coordinates for each location ID, such as both (x1,y1) and (x2,y2). The location ID can then be defined to be included in the circular scanning window either (i) if at least one of the coordinate sets is located within the circle, or (ii) if and only if all of the coordinate sets are within the circle. The multiple sets of coordinates are specified in the coordinates file, with each one on a separate row.

**Related Topics:** *Advanced Features, ASCII File Format, Input Tab, Meta Location File, Special Neighbors File.*

---

## Spatial Window Tab



*Spatial Window Tab Dialog Box*

Use the Spatial Window Tab to define the exact nature of the scanning window with respect to space.

**Related Topics:** *Advanced Features, Analysis Tab, Temporal Window Tab, Maximum Spatial Cluster Size, Include Purely Temporal Clusters.*

### Maximum Spatial Cluster Size

The program will scan for clusters of geographic size between zero and some upper limit defined by the user. The upper limit can be specified either as a percent of the population used in the analysis, as a percent of some other population defined in a max circle size file, or in terms of geographical size using the circle radius. The maximum can also be defined using a combination of these three criteria.

The recommended choice is to specify the upper limit as a percent of the population at risk, and to use 50% as the value. It is possible to specify a maximum that is less than 50%, but not more than 50%. A cluster of larger size would indicate areas of exceptionally low rates outside the circle rather than an area of exceptionally high rate within the circle (or vice-versa when looking for clusters of low rates). When in doubt, choose a high percentage, since SaTScan will then look for clusters of both small and large sizes without any pre-selection bias in terms of the cluster size. When calculating the percentage, SaTScan uses the population defined by the cases and controls for the Bernoulli model, the covariate adjusted population at risk from the population file for the discrete Poisson model, the cases for the space-time permutation, multinomial, ordinal, exponential and normal models and the size of the circle as

a percentage of the total area in the polygons for the continuous Poisson model. When there are multiple data sets, the maximum is defined as a percentage of the combined total population/cases in all data sets.

It is also possible to specify the maximum circle size in terms of actual geographical size rather than population. If latitude/longitude coordinates are used, then the maximum radius should be specified in kilometers. If Cartesian coordinates are used, the maximum radius should be specified in the same units as the Cartesian coordinates.

Alternatively, for the discrete scan statistics, it is possible to specify a max circle size file to define the maximum circle size. This file must contain a 'population' for each location, and the maximum circle size is then defined as a percentage of this population rather than the regular one. This feature may be used when, for example, you want to define the circles in the Bernoulli or space-time population models based on the actual population rather than the locations of cases and controls. It may also be used if you want the geographical circles to include for example at most 10 counties out of a total of 100, irrespectively of the population in those counties. This is accomplished by assigning a 'population' of 1 to each county in the special max circle size file and then set the maximum circle size to be 10% of this 'population'.

If a prospective space-time analysis is performed, adjusting for earlier analyses, and if the max circle size is defined as a percentage of the population, then the special max circle size file must be used. This is to ensure that the evaluated geographical circles do not change over time.

**Related Topics:** *Advanced Features, Spatial Window Tab, Max Circle Size File, Include Purely Temporal Clusters, Computing Time.*

## **Include Purely Temporal Clusters**

A purely temporal cluster is one that includes the whole geographic area but only a limited time period. When doing a space-time analysis, it is possible to allow potential clusters to contain the whole geographical area under study, as an exception to the maximum spatial cluster size chosen. In this way, purely temporal clusters are included among the collection of windows evaluated.

**Note:** This option is not available for the space-time permutation model, as that model automatically adjusts for purely temporal clusters. When adjusting for purely temporal clusters using stratified randomization, all purely temporal clusters are adjusted away, and this parameter has no effect on the analysis.

**Related Topics:** *Advanced Features, Spatial Window Tab, Maximum Spatial Cluster Size, Include Purely Spatial Clusters, Temporal Trend Adjustment, Computing Time.*

## **Elliptic Scanning Window**

As an advanced option, it is possible to use a scanning window that consists not only of circles but also of ellipses of different shapes and angles. When the elliptic spatial scan statistic is requested, SaTScan uses the circular window plus five different elliptic shapes where the ratio of the longest to the shortest axis of the ellipse is 1.5, 2, 3, 4 or 5. For each shape, a different number of angles of the ellipse are used, to the number being 4, 6, 9, 12 and 15 respectively, depending on the elliptic shape. The north-south axis is always one of the angles included, and the remainder is equally spaced around the circle. For each

shape and angle, all possible sizes of the ellipses are used, up to an upper limit specified by the user in the same way as for the circular window.

When using an elliptic window shape, it is possible to request a non-compactness (eccentricity) penalty, which will favor more compact over less compact ellipses even when they have slight lower likelihood ratios but the less compact ellipses when the difference is larger. The formula for the penalty is  $[4s/(s+1)^2]^a$ , where  $s$  is the elliptic window shape defined as the ratio of the length of the longest to the shortest axis of the ellipse. With a strong penalty  $a=1$ , with a medium penalty  $a=1/2$  and with no penalty  $a=0$ .

**Note:** In batch mode, it is possible to request SaTScan to use any other collection of ellipses to define the scanning window and any value of the eccentricity penalty parameter greater than zero.

**Note:** The elliptic window option can only be used when regular two-dimensional Cartesian coordinates are used, but not when they are specified as latitude/longitude. If you have the latter, you must first do a planar map projection from the latitude/longitude coordinates, of which there are many different ones proposed in the geography literature.

**Note:** The elliptic scanning window is not available for the continuous Poisson model.

**Related Topics:** *Advanced Features, Computing Time, Include Purely Spatial Clusters, Likelihood Ratio Test, Maximum Spatial Cluster Size, Spatial Temporal and Space-Time Scan Statistics, Spatial Window Tab.*

## Isotonic Spatial Scan Statistic

In the standard spatial scan statistic, the alternative model is that there is a higher rate inside the cluster and a lower rate outside the cluster. While the rate is not assumed to be constant neither inside nor outside the window, any such variation in the rate is not taken into account when calculating the likelihood. Rather than one circular window, the isotonic spatial scan statistic<sup>13</sup> defines the window by using a set of overlapping circles of different size that are centered on the same point. The alternative model is that the rate is highest within the innermost circle, somewhat lower between the first and second circles, and so on, until the last circle. There is no predefined number of circles or any prior assumption on their respective size, except that the biggest circle must be smaller than the user specified max circle size. Rather, the method finds the collection of circles that maximizes the likelihood ratio statistic. As with the standard spatial scan statistic, the window moves over space, considering many possible circle centroids. The method adjusts for the multiple testing inherent in both the many cluster locations considered as well as the many possible collections of circles used for the scanning window.

**Note:** While the method evaluates a window with multiple circles, it is sometimes a single circle that provides the highest likelihood and therefore defines the most likely cluster.

**Note:** The isotonic spatial scan statistic is only available for purely spatial analyses using either the discrete Poisson or the Bernoulli models. It is not available for multiple data sets.

**Related Topics:** *Advanced Features, Maximum Spatial Cluster Size.*

---

## Temporal Window Tab

The screenshot shows the 'Advanced Analysis Features' dialog box with the 'Temporal Window' tab selected. The 'Maximum Temporal Cluster Size' section has two radio buttons: 'is 50.0 percent of the study period (<= 90%, default = 50%)' (selected) and 'is 1 years'. Below this is a checkbox for 'Include Purely Spatial Clusters (Temporal Size = 100%)'. The 'Flexible Temporal Window Definition' section has a checkbox for 'Include only windows with:' followed by two rows of date pickers. The first row is 'Start time in range: 2000 1 1 to 2000 12 31' and the second row is 'End time in range: 2000 1 1 to 2000 12 31'. At the bottom right are 'Set Defaults' and 'Close' buttons.

***Temporal Window Tab Dialog Box***

Use the Temporal Window Tab to define the exact nature of the scanning window with respect to time.

**Related Topics:** *Advanced Features, Analysis Tab, Spatial Window Tab, Maximum Temporal Cluster Size, Include Purely Spatial Clusters, Flexible Temporal Window Definition.*

### Maximum Temporal Cluster Size

For purely temporal and space-time analyses, the maximum temporal cluster size can be specified in terms of a percentage of the study period as a whole or as a certain number days, months or years. The maximum must be at least as large as the length of aggregated time interval length. If specified as a percent, then for the Bernoulli and Poisson models, it can be at most 90 percent, and for the space-time permutation model, at most 50 percent. The recommended value is 50 percent

**Related Topics:** *Temporal Window Tab, Maximum Spatial Cluster Size, Include Purely Spatial Clusters, Flexible Temporal Window Definition, Time Aggregation.*

### Include Purely Spatial Clusters

In addition to the maximum temporal cluster size, it is also possible to allow clusters to contain the whole time period under study. In this way, purely spatial clusters are included among the evaluated windows. The purpose of specifying a maximum temporal size, but still including purely spatial clusters, is to

eliminate clusters containing the whole study period except a small time period at the very beginning or at the very end of the study period.

**Note:** When adjusting for purely spatial clusters using stratified randomization, all purely spatial clusters are adjusted away, and this parameter has no effect on the analysis.

**Related Topics:** *Temporal Window Tab, Maximum Temporal Cluster Size, Include Purely Temporal Clusters, Spatial Adjustment.*

## **Flexible Temporal Window Definition**

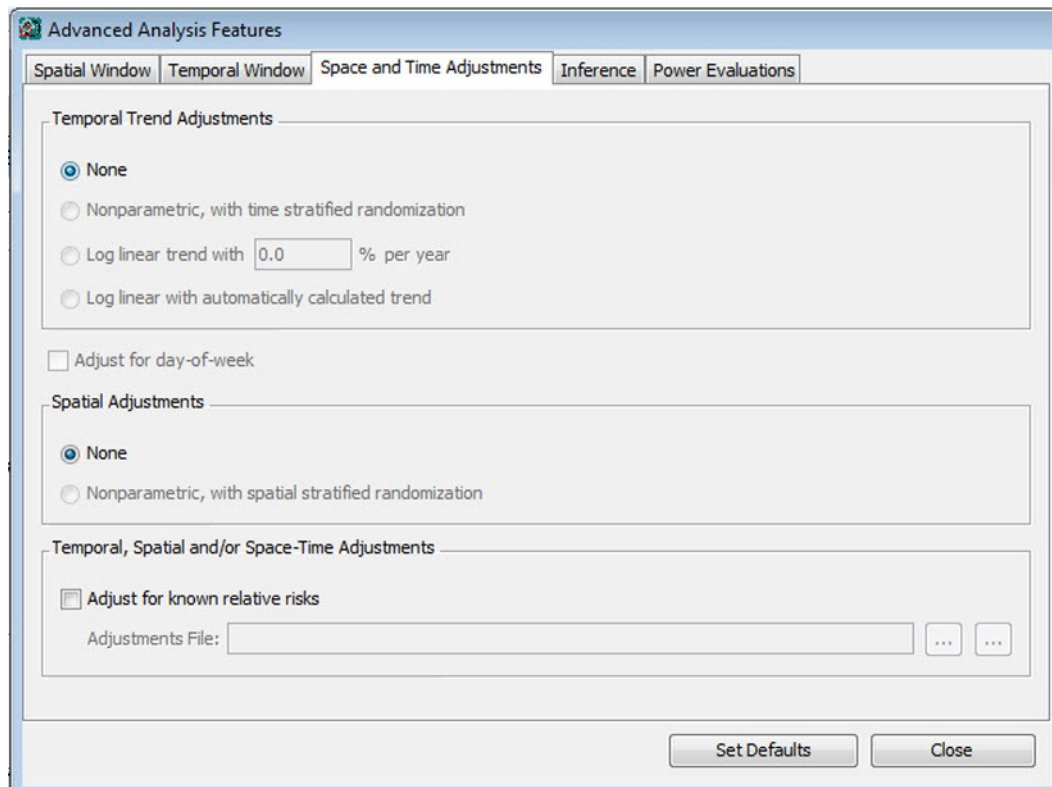
For retrospective analyses, SaTScan will evaluate all temporal windows less than the specified maximum, and for prospective analyses the same is true with the added restriction that the end of the window is identical to the study period end date. When needed, SaTScan can be more flexible than that, and it is possible to define the scanning window as any time period that start within a predefined ‘start range’ and ends within a predefined ‘end range’.

This option is only available when a retrospective purely temporal or a retrospective space-time analysis is selected on the Analysis Tab.

**Related Topics:** *Temporal Window Tab, Maximum Temporal Cluster Size, Include Purely Spatial Clusters, Study Period, Time Aggregation.*

---

## Spatial and Temporal Adjustments Tab



*Spatial and Temporal Adjustments Tab Dialog Box*

Covariates are adjusted for either by including them in the case and population files or by using multiple data sets, depending on the probability model used. The features on this tab are used to adjust for temporal, spatial and space-time trends and variation. Most are only available when using the discrete Poisson probability model. The one exception is the space by day-of-week interaction adjustment for the space-time permutation model.

**Related Topics:** *Advanced Features, Analysis Tab, Spatial and Temporal Adjustments, Temporal Trend Adjustment, Spatial Adjustment, Adjustment with Known Relative Risk, Poisson Model.*

### Temporal Trend Adjustment

Temporal trends can be adjusted for in three different ways:

*Non-parametric:* When the adjustment is non-parametric, SaTScan adjusts for any type of purely temporal variation. This is done by stratifying the randomization by the aggregated time intervals, so that each time interval has the same number of cases in the real and random data sets. That is, it is only the spatial location of a case that is randomized.

*Log linear trend, specified by user:* Specify an annual percent increase or decrease in the risk. A decreasing trend is specified with a negative number. For example, if the rate decreases by 1.4 percent per year, then write "-1.4" in the "% per year" box.

*Log linear trend, automatically calculated:* Rather than the user specifying the adjusted relative risk, SaTScan can calculate the observed trend in the data and then adjust for exactly that amount of increase or decrease.

The default is no temporal trend adjustment.

**Note:** When doing a spatial variation in temporal trends analysis, it is not possible to adjust for temporal trends.

**Related Topics:** *Spatial and Temporal Adjustment Tab, Day-of-Week Adjustment, Spatial and Temporal Adjustments, Spatial Adjustment, Adjustment with Known Relative Risk, Poisson Model.*

## Day-of-Week Adjustment

For some data, such as physician visits, there may be natural day-of-week variation that should be adjusted for. This is done in a non-parametric fashion. The effect of requesting a day-of-week adjustment on the Space and Time Adjustments Tab is the same as the effect of including a day-of-week covariate in the input files.

With the space-time permutation model, day-of-week is automatically adjusted for whether specifically requested or not, as part of its complete adjustment for any purely temporal variation. Instead, it is possible to request an adjustment for day-of-week by space interaction. This adjusts for the fact that some geographical areas may have a different day-of-week pattern than other areas. For example, one medical clinic may have a large number of weekend visits while another clinic may be closed on weekends. The effect is the same as including day-of-week as a covariate in the case file used by the space-time permutation model.

**Related Topics:** *Spatial and Temporal Adjustment Tab, Temporal Trend Adjustments, Adjustment with Known Relative Risk, Space-Time Permutation Model.*

## Spatial Adjustment

When a purely spatial analysis is performed the purpose is to find purely spatial clusters. For the space-time scan statistic, this feature adjusts away all such clusters, to see if there are any space-time clusters not explained by purely spatial clusters. This is done in a non-parametric fashion, through stratified randomization by location, so that the total number of cases in each specific location is the same in the real and random data sets. That is, only the time of a case is randomized.

The default is no spatial adjustment.

**Note:** It is not possible to simultaneously adjust for spatial clusters and purely temporal clusters using stratified randomization. If both types of adjustments are desired, the space-time permutation model should be used instead. It is possible to adjust for purely spatial clusters with stratified randomization together with a temporal adjustment using a log linear trend.

**Related Topics:** *Spatial and Temporal Adjustment Tab, Spatial and Temporal Adjustments, Temporal Trend Adjustment, Adjustment with Known Relative Risk, Poisson Model.*



## Adjustment with Known Relative Risks

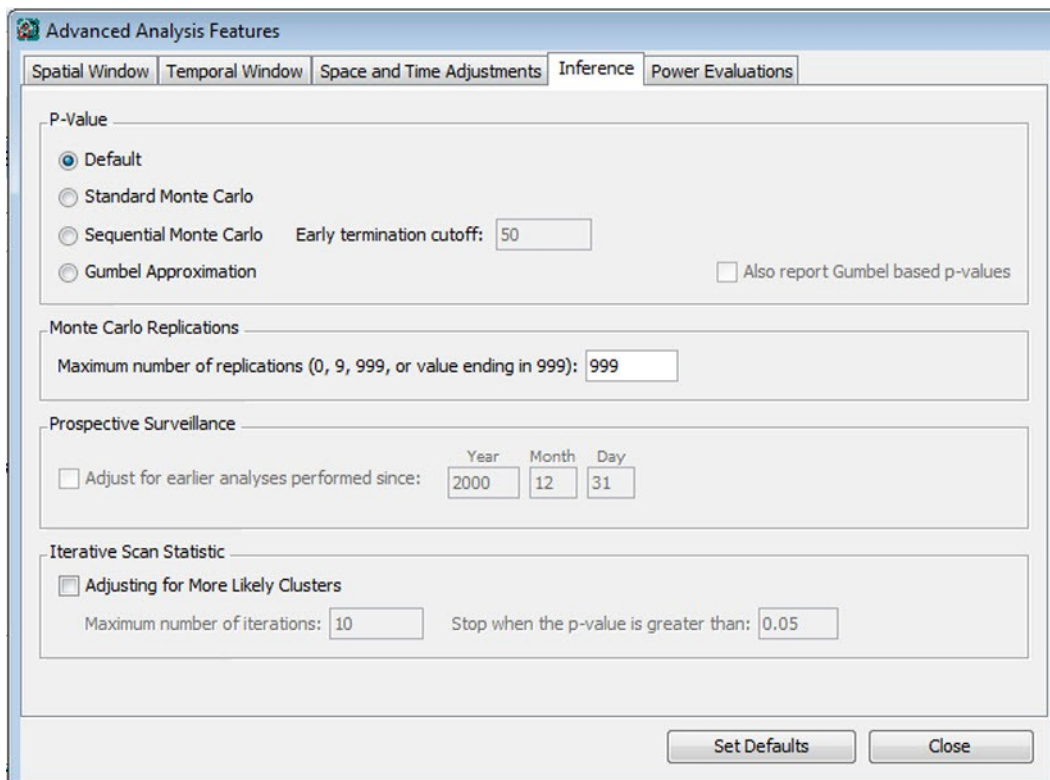
The most flexible way to adjust a discrete Poisson model analysis is to use the special adjustments file. In this file, a relative risk is specified for any location and time period combination, and SaTScan will adjust the expected counts up or down based on this relative risk. One use of this option is to adjust for missing data, by specifying a zero relative risk for those location and time combinations for which data is missing.

The required format of the Adjustments File is described in the section on Input data.

**Related Topics:** *Spatial and Temporal Adjustment Tab, Spatial and Temporal Adjustments, Temporal Trend Adjustment, Spatial Adjustment, Adjustments File, Poisson Model.*

---

## Inference Tab



The screenshot shows the 'Advanced Analysis Features' dialog box with the 'Inference' tab selected. The dialog has five tabs: 'Spatial Window', 'Temporal Window', 'Space and Time Adjustments', 'Inference', and 'Power Evaluations'. The 'Inference' tab contains the following settings:

- P-Value:** Four radio buttons are present: 'Default' (selected), 'Standard Monte Carlo', 'Sequential Monte Carlo', and 'Gumbel Approximation'. To the right of the 'Sequential Monte Carlo' option is a text box labeled 'Early termination cutoff:' with the value '50'. To the right of the 'Gumbel Approximation' option is a checkbox labeled 'Also report Gumbel based p-values' which is unchecked.
- Monte Carlo Replications:** A text box labeled 'Maximum number of replications (0, 9, 999, or value ending in 999):' contains the value '999'.
- Prospective Surveillance:** A checkbox labeled 'Adjust for earlier analyses performed since:' is unchecked. To its right are three text boxes for 'Year' (2000), 'Month' (12), and 'Day' (31).
- Iterative Scan Statistic:** A checkbox labeled 'Adjusting for More Likely Clusters' is unchecked. Below it are two text boxes: 'Maximum number of iterations:' with the value '10', and 'Stop when the p-value is greater than:' with the value '0.05'.

At the bottom right of the dialog are two buttons: 'Set Defaults' and 'Close'.

*Inference Tab Dialog Box*

This tab is reached by clicking the Advanced button in the lower right corner of the Analysis Tab.

**Related Topics:** *Advanced Features, Analysis Tab, Early Termination of Simulations, Adjust for Earlier Analyses in Prospective Surveillance.*

## P-Value

To calculate p-values for detected clusters, SaTScan program uses computer simulations to generate a number of random replications of the data set under the null hypothesis. If the maximum likelihood ratio calculated for the most likely cluster in the real data set is high compared to the maximum likelihood ratios calculated for the most likely clusters in the random data sets, that is evidence against the null hypothesis and for the existence of clusters. The comparison can be done in either of three ways, or by using a combination of them, the latter being the default option.

*Standard Monte Carlo:* The test statistic is calculated for each random replication as well as for the real data set, and if the latter is among the 5 percent highest, then the test is significant at the 0.05 level. If it is among the 1 percent highest, the test is significant at the 0.01 level, and so on. This is called Monte Carlo hypothesis testing, and was first proposed by Dwass<sup>15</sup>. Irrespective of the number of Monte Carlo replications chosen, the hypothesis test is unbiased, resulting in a correct significance level that is neither conservative nor liberal nor an estimate. The number of replications does affect the power of the test, with more replications giving slightly higher power.

In SaTScan, the number of replications must be at least 999 to ensure excellent power for all types of data sets. For small to medium size data sets, 9999 replications are recommended since computing time is not a major issue.

*Sequential Monte Carlo:* With more Monte Carlo replications, the power of the scan statistic is higher, but it is also more time consuming to run. When the p-value is small, this is often worth the effort, but for large p-values it is often irrelevant whether for example  $p=0.7535$  or  $p=0.8545$ . SaTScan provides the option to terminate the Monte Carlo simulations early when the p-value is large, by employing the sequential Monte Carlo test.<sup>16, 17</sup> With this option, the SaTScan calculations will terminate as soon as a fixed number of Monte Carlo replicas has a likelihood ratio that is larger than the likelihood ratio from the real data set. The default value is 50 replicas. If the fixed number is never reached, the calculations will continue until the maximum number of Monte Carlo replicas has been reached. With the default values of 50 and 999, there is no loss of power at the  $\alpha=0.05$  level, when comparing the sequential to the standard Monte Carlo test.

*Gumbel Approximation:* With 999 random replicas, the lowest p-value that the Monte Carlo hypothesis testing can report is  $1/(999+1)=0.001$ . Likewise, with 9999 replicas, the lowest possible p-value is 0.0001. As an alternative option to Monte Carlo hypothesis testing, it is possible to employ the Gumbel extreme value distribution to estimate approximate p-values.<sup>18</sup> With this approach, there is no lower limit on the resulting p-values (other than  $p>0$ , of course). The method works by first generating 999, or some other number of random replicas of the data under the null hypothesis. The maximum likelihood ratio from each replica is then used to fit a Gumbel distribution to the data using methods of moments estimation. Once the Gumbel distribution that best fits the data has been obtained, the p-value is calculated as the probability that this distribution generates a value greater than the maximum likelihood ratio observed for the most likely cluster from the real data set.

For the purely spatial scan statistic with the discrete Poisson and Bernoulli probability models, it has been shown that the Gumbel distribution fits the data very well and that it generates very accurate p-values<sup>18</sup>. There have not yet been any similar studies for the other scan statistics, so this option is for the time being only available for purely spatial analyses with those two probability models.

*Default P-value:* As the default, SaTScan will calculate the p-values by using a combination of the three manners described above. For example, it may present the sequential Monte Carlo based p-value unless the p-value is very small, in which case it will report the Gumbel approximation. The exact approach depends on the type of analysis requested and the nature of the data, since the sequential Monte Carlo and the Gumbel approximation does not work for all analyses and data sets

**Note:** In prior versions of SaTScan, the standard Monte Carlo was the default method for calculating p-values, and the other options are new to version 9.0.

**Related Topics:** *Analysis Tab, Computing Time, Inference Tab, Likelihood Ratio Test, Monte Carlo Rank, Monte Carlo Replications, Random Number Generator, Results of Analysis.*

## **Adjust for Earlier Analyses in Prospective Surveillance**

When doing prospective purely temporal or prospective space-time analyses repeatedly in a time-periodic fashion, it is possible to adjust the statistical inference (p-values) for the multiple testing inherent in the repeated analyses done. To do this, simply check the “adjust for earlier analyses” box, and specify the date for which you want to adjust for all subsequent analyses. This date must be greater or equal to the study period start date and less than or equal to the study period end date, as specified on the Input Tab.

The adjustment is done by using a different set of cylinders when calculating the maximum likelihood for the real and random data sets. For the real data set, the maximum is obtained over all currently alive clusters, that is, those cylinders that reach the study period end date. For the random data sets, the maximum is taken over all cylinders with an end date after the adjustment date specified on this date. That is, it takes the maximum over all cylinders previously used in the prior analyses that are now being adjusted for.

For the adjustment to be correct, it is important that the scanning spatial window is the same for each analysis that is performed over time. This means that the grid points defining the circle centroids must remain the same. If the location IDs in the coordinates file remain the same in each time-periodic analysis, then there is no problem. On the other hand, if new IDs are added to the coordinates file over time, then you must use a special grid file and retain this file through all the analyses. Also, when you adjust for earlier analyses, and if the max circle size is defined as a percentage of the population, then the special max circle size file must be used.

**Related Topics:** *Inference Tab, Computing Time, Type of Analysis, Spatial Temporal and Space-Time Scan Statistics.*

## **Iterative Scan Statistic**

The iterative scan option is used to adjust the p-values of secondary clusters for more likely clusters that are found and reported. This is done by doing the analysis in several iterations, removing the most likely cluster found in each iteration, and then reanalyzing the remaining data<sup>23</sup>. The user must specify the maximum number of iterations allowed, in the range 1-32000. The user may also request that the iterations stop when the last found cluster has a p-value greater than a specified lower bound.

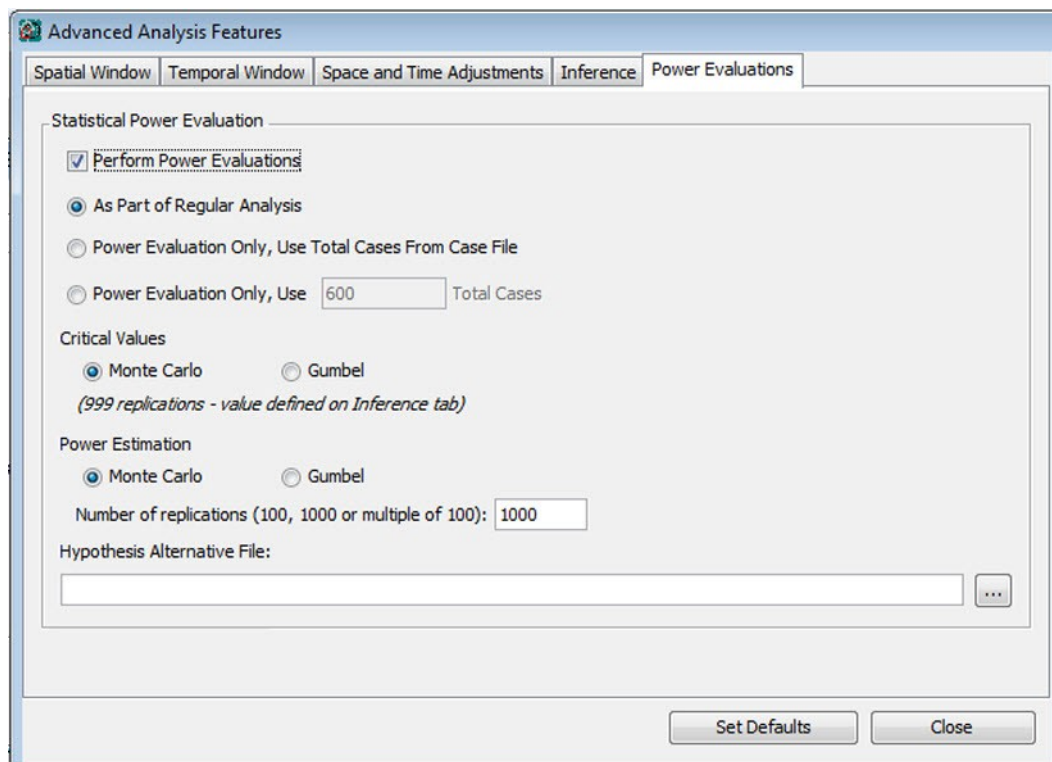
In terms of computing time, each iteration takes approximately the same amount of time as a regular analysis with the same parameters.

**Note:** It has been shown that the iterative scan statistic p-values are valid for a purely spatial analysis with the discrete Poisson model. The feature is also available for the other discrete scan statistics, but it is not known whether the p-values are as accurate. The iterative scan statistic is not available for space-time scan statistics, or for the continuous Poisson model.

**Related Topics:** *Adjusting for More Likely Clusters, Inference Tab, Computing Time.*

---

## Power Estimation Tab



*Power Estimation Tab Dialog Box*

Just as there is no known analytical way to obtain the p-values for the scan statistics in the SaTScan software, there is no analytical way to estimate the power. Hence, that needs to be done through simulations. By specifying a set of hypothetical cluster locations, sizes and relative risks, SaTScan will estimate the statistical power for those pre-specified alternative hypotheses. Note that several alternative hypotheses can be evaluated at the same time, which is computationally more efficient than evaluating one at a time. The alternative hypotheses are defined in the Alternative Hypothesis File, with an empty row between the different alternative hypotheses of interest. Note that all locations in the same cluster does not have to have the same relative risk, and it is even okay to have some relative risks greater than one and other relative risks less than one.

Power estimations can be done either in connection with an actual SaTScan analysis of real data, or as a separate analysis. For the latter, it is not necessary to specify a case file, but sufficient to specify the total number of cases.

The precision of the power estimates will increase with more simulation runs. We recommend at least 9999 simulations under the null hypothesis and at least 1000 under each of the alternative hypotheses. Power estimations can be conducted using either a straight Monte Carlo approach or by using the Gumbel approximation. The results should be similar.

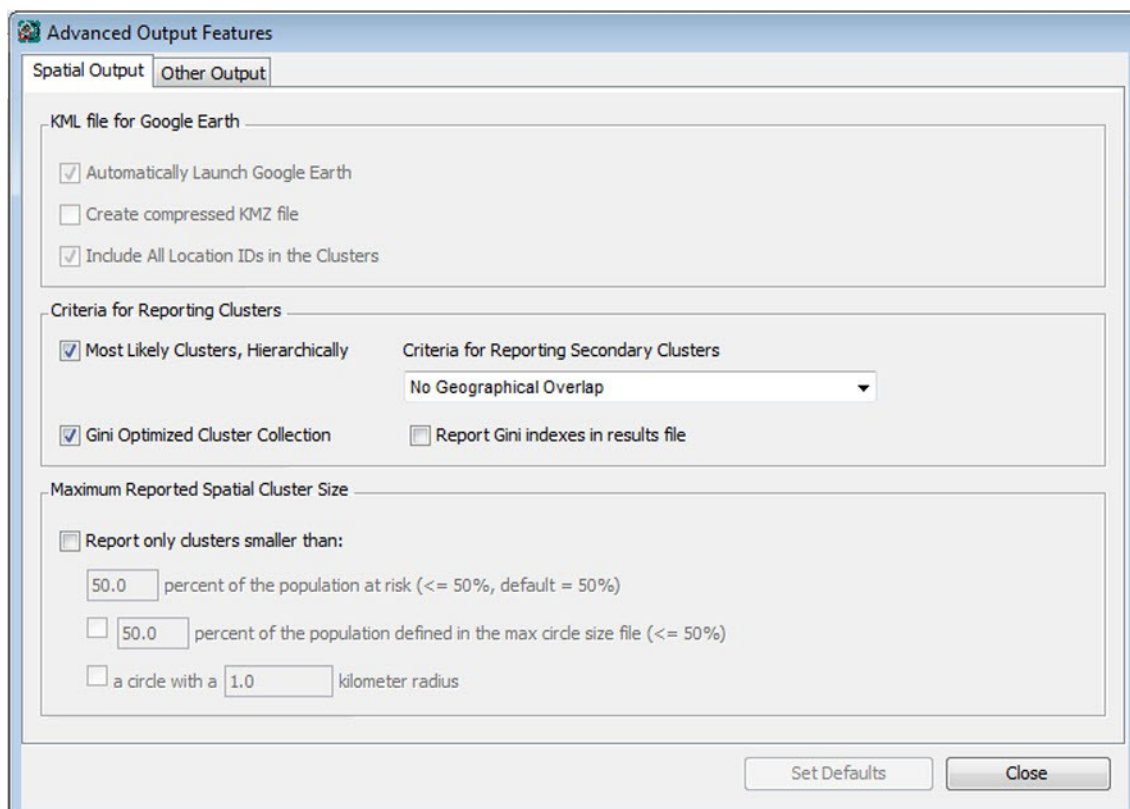
The statistical power will vary greatly depending on the total number of cases in the data set, on the population size of the cluster and on the relative risk in the cluster, with higher values resulting in higher power.

This feature is only available for the discrete Poisson probability model

**Related Topics:** *Poisson Probability Model, Alternative Hypothesis File*

---

## Spatial Output Tab



***Spatial Output Tab Dialog Box***

This tab is reached by clicking the Advanced button in the lower right corner of the Output Tab.

**Related Topics:** *Advanced Features, Output Tab, Results of Analysis, Criteria for Reporting Secondary Clusters, Report Only Small Clusters.*

## KML Geographical Output File (\*.kml)

With the KML output file, the detected SaTScan clusters can be depicted using Google Earth or Google Maps. Google Earth is free software that can be downloaded from 'earth.google.com'. If it is on the computer, SaTScan will automatically launch it after the analysis is completed and show the detected clusters. The Google Earth map can then be edited to add or subtract other features to the map, such as country boundaries, city name labels, railroads and volcanoes.

The KML output file can only be used when the geographical coordinates are specified using latitudes and longitudes. While KML files can also be read by some other geographical software, the SaTScan KML files may or may not work well for those applications.

**Related Topics:** *Output Tab, Standard Results File, Shapefile Geographical Output, Spatial Output Tab, Cluster Information File, Geographical Coordinates File, Lat/Long Coordinates, Column Output Format.*

## Criteria for Reporting Secondary Clusters

SaTScan evaluates an enormous amount of different circles/cylinders in order to find the most likely cluster. For large data sets the number of potential clusters is in the millions for a purely spatial analysis and in the billions for a space-time analysis. All of these other clusters may be considered secondary clusters with either a high or a low rate. To present all of these secondary clusters is impractical and unnecessary since many of them will be very similar to each other. For example, to add one location with a very small population to the most likely cluster will not decrease the likelihood very much, even if that location contains no additional cases. Such a secondary cluster is not interesting even though it could have the second highest likelihood among all the clusters evaluated.

Rather than reporting information about all evaluated clusters, SaTScan only reports a limit number of secondary clusters using criteria specified by the user. A three-stage procedure is used to select the secondary clusters to report:

1. For each circle centroid, SaTScan will only consider the cluster with the highest likelihood among those that share that same centroid (grid point).
2. The resulting set of clusters will be ordered in descending order by the value of their log likelihood ratios, creating a list with the same number of clusters as there are grid points.
3. The most likely cluster will always be reported. Secondary clusters can be reported based in either or both of two criteria. The first is a purely hierarchical criterion where secondary clusters with  $p < 1.0$  are reported depending on their degree of geographical overlap with more likely clusters already reported. The second is a criteria based on the Gini index, which is a measure of statistical dispersion. With this criterion, SaTScan selected the group of non-overlapping clusters that maximizes the Gini index, so that there is a big difference in rates between the cluster and non-cluster areas<sup>31</sup>. These latter are called 'gini clusters'.

## Hierarchical Cluster Reporting Options

**No Geographical Overlap:** Default. Secondary clusters will only be reported if and only if they do not overlap with a previously reported cluster, that is, they may not have any location IDs in common.

Therefore, no overlapping clusters will be reported. This is the most restrictive option, presenting the fewest number of clusters.

**No Cluster Centers in Other Clusters:** Secondary clusters are reported if they are not centered in a previously reported cluster and do not contain the center of a previously reported cluster. While two clusters may overlap, there will be no reported cluster with its centroid contained in another reported cluster.

**No Cluster Centers in More Likely Clusters:** Secondary clusters are reported if they are not centered in a previously reported cluster. This means that there will be no reported cluster with its center contained in a previously reported more likely cluster.

**No Cluster Centers in Less Likely Clusters:** Secondary clusters do not contain the center of a previously reported cluster. This means that there will be no reported cluster with its center contained in a subsequently reported less likely cluster.

**No Pairs of Centers Both in Each Others Clusters:** Secondary clusters are not centered in a previously reported cluster that contains the center of a previously reported cluster. This means that there will be no pair of reported clusters each of which contain the center of the other.

**No Restrictions = Most Likely Cluster for Each Grid Point:** The most extensive option is to all present clusters in the list, with no restrictions. This option reports the most likely cluster for each grid point, including clusters with  $p=1.0$ . This means that the number of clusters reported is identical to the number of grid points. WARNING: This option may create output files that are very large in size.

### **Gini Index Cluster Reporting Option**

This option is only valid for purely spatial analyses with the Poisson or Bernoulli models. To create the collection of clusters based on the Gini index, SaTScan first defines a collection of upper limits on the cluster size, with the default collection being 1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 25, 30, 35, 40, 45 and 50 percent of the population. For each upper limit, the hierarchical no-geographical overlap cluster collection criterion is used to define a set of clusters, as described above. The Gini index is then calculated for this set of clusters, and when repeated for each upper limit, we get twelve different Gini indices. SaTScan then picks the collection that maximizes the Gini index. These are called ‘Gini clusters’. As an optional feature, SaTScan will report the values of the Gini indices for each of the upper limits used.

**Note:** It is possible to request that both the hierarchical and Gini clusters are reported, and this is the default setting. This means that there may be overlapping clusters even when the hierarchical non-overlapping cluster option is selected, since some of the hierarchical clusters may overlap with some of the Gini clusters, without being identical. If none are requested, SaTScan will only report the most likely cluster.

**Note:** The criteria for determining overlap is based only on geography, ignoring time. Hence, in a space-time analysis, a secondary cluster may not be reported if it is in the same location as a more likely cluster, even if they are non-overlapping in time.

**Related Topics:** *Advanced Features, Inference Tab, Results of Analysis, Maximum Spatial Cluster Size, Report Only Small Clusters.*

## Maximum Reported Spatial Cluster Size

The maximum spatial cluster size is specified on the Analysis > Advanced > Spatial Window tab. The default is that SaTScan will report the most likely cluster among these as well as any secondary clusters. This option allows you to specify a different maximum size on the cluster that are evaluated and those that are reported. The latter maximum must be smaller though.

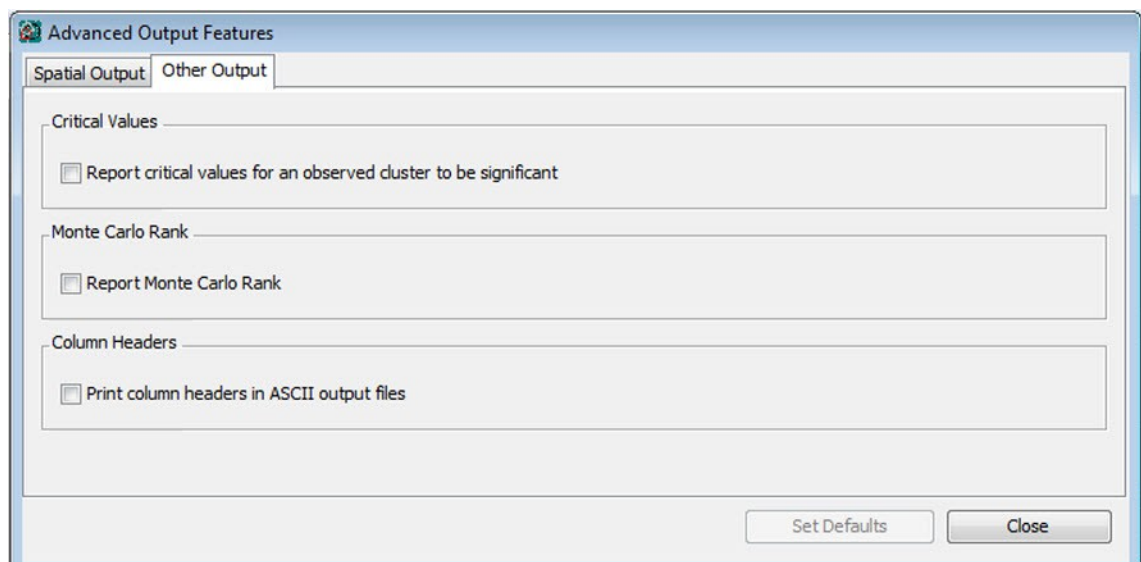
It is natural to ask why anyone would want to specify a different maximum reported cluster size than simply changing the maximum size of the clusters being evaluated. The reason is as follows. When the most likely cluster is very large in size, it is sometimes of interest to know whether it contains smaller clusters that are statistically significant on their own strength. One way to find such clusters is to play around with the maximum spatial circle size parameter on the Spatial Window tab, but that leads to incorrect statistical inference as the maximum size on the circles evaluated is then chosen based on the results of the analysis, leading to pre-selection bias. To avoid this problem, this option allows you to keep the original maximum spatial circle size of the clusters that SaTScan evaluates and uses for the statistical inference, and at the same time limit the size of the clusters that are reported. This means that the p-values for the smaller reported clusters are adjusted for all size clusters including those that were not allowed to be reported. That is, SaTScan reports clusters based on this second maximum but it adjusts for the multiple testing inherent in the larger collection of circles defined by the first maximum.

The unit by which to define the maximum size of the reported cluster is the same as the unit used to define the maximum cluster size for inference purposes, as defined on the Spatial Window Tab.

**Related Topics:** *Advanced Features, Inference Tab, Results of Analysis, Criteria for Reporting Secondary Clusters, Maximum Spatial Cluster Size, Log Likelihood Ratio, Standard Results File.*

---

## Other Output Tab



*Other Output Tab Dialog Box*



## Critical Values

When selecting this option, SaTScan will report the critical values needed in order for a cluster to be statistically significant at the 0.05 and 0.01 alpha levels. The critical values are reported on the standard results file.

**Related Topics:** *Standard Results File.*

## Monte Carlo Rank

When using standard or sequential Monte Carlo testing, this option will in addition to the p-value also report the rank of the log likelihood ratio from the most likely cluster in the real data set among all the maximum log likelihood ratios from the random data sets replicated under the null hypothesis.

**Related Topics:** *P-Value, Standard Results File.*

## Column Headers


Check this box if you want column headers in the Other Output files.

**Related Topics:** *Column Output Format, Column Information File, Location Information.*

# Running SaTScan

---

## Specifying Analysis and Data Options

The SaTScan program requires that you specify parameters defining input, analysis and output options for the analysis you wish to conduct. A tabbed dialog is provided for this purpose. To access the parameter tab dialog, either press the  button or select the File/New menu item. Specify the parameters for your session on the following tabs:

- Input Tab
- Analysis Tab
- Output Tab

See the section on Basic SaTScan Features for instructions on how to fill in these tabs.


Most analyses can be performed using only these three tabs. For each tab, there are additional features that can be selected by first clicking on the Advanced button in the lower right corner of the tab. These additional features may be useful in special circumstances.

The available choices for some features may depend on what was selected in other places. For example, if a purely spatial analysis is chosen, the space-time permutation model is not available, and vice versa.

**Related Topics:** *Basic SaTScan Features, Input Tab, Analysis Tab, Output Tab, Advanced Features, Launching the Analysis.*

---

## Launching the Analysis

Once the data input files have been created, and the parameters defining the input, analysis and output options have been specified, select the Execute  button to launch the analysis and produce the results file. A special job status window will appear containing status, warning and/or error messages. Once the analysis has been completed, the standard results file will appear in the job status window.

Multiple parameter session windows may be opened simultaneously for data entry, and multiple analyses may be run concurrently. If you are running multiple analyses concurrently, please verify that the output files have different names.

**Related Topics:** *Input Data, Data Requirements, Specifying Analysis and Data Options, Status Messages, Warnings and Errors, Computing Time, Batch Mode.*

---

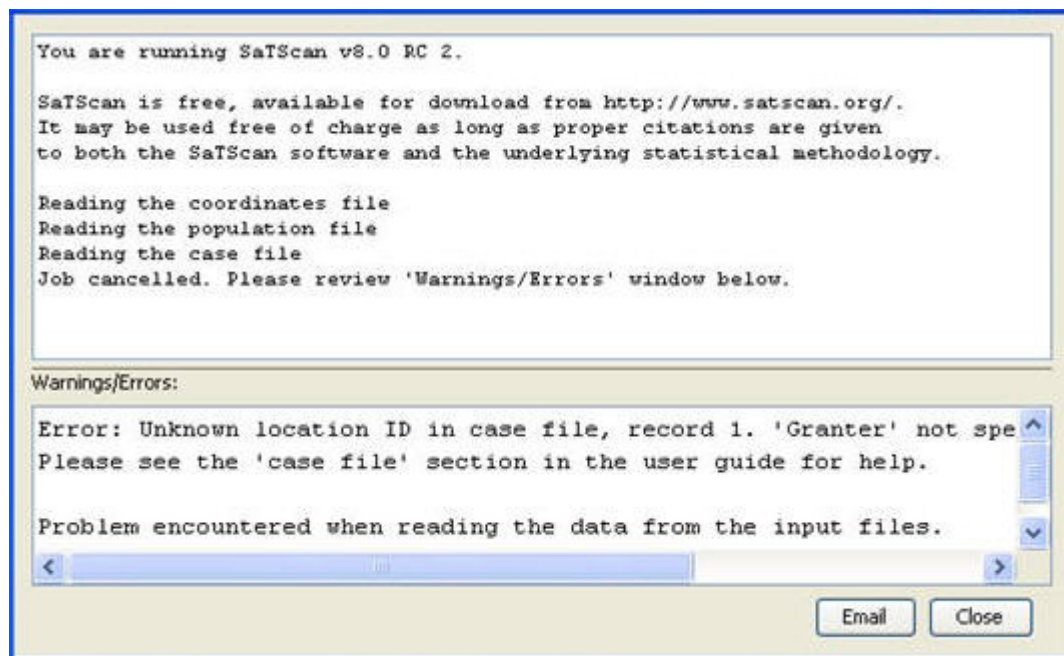
## Status Messages

Status messages are displayed as the program executes the analysis, as the data is read, and at each step of the analysis. Normal status messages are displayed in the top box of the job status window. Warnings and error messages are displayed in the bottom box of the job status window. Upon successful completion of the calculations, the standard results file will be shown in the job status window.

**Related Topics:** *Launching the Analysis, Warnings and Errors.*

---

## Warnings and Errors



*SaTScan Status Messages and Warnings/Errors Dialog Box*

### *Warning Messages*

SaTScan may produce warnings as the job is executing. If a warning occurs, a message is displayed in the Warnings/Errors box on the bottom of the job status window. A warning will not stop the execution of the analysis. If a warning occurs, please review the message and access the help system if further information is required.

If you do not want to see the warning messages, they can be turned off by clicking “Session > Execute Options > Do not report warning messages”.

### *Error Messages*

If a serious problem occurs during the run, an error message will be displayed in the Warnings/Errors box on the bottom of the job status window and the job will be terminated. The user may resolve most errors by reviewing the message and using the help system.

One of the most common errors is that the input files are not in the required format, or that the file contents are incompatible with each other. When this occurs, an error message will be shown specifying the nature and location of the problem. Such error messages are designed to help with data cleaning.

If the error message cannot be resolved, you may press the email button on the job status window. This will generate an automatic email message to SaTScan technical support. The contents of the “Warnings/Errors” box will be automatically placed in the e-mail message. All a user needs to do is press their e-mail Send key. Users may also print the contents of the Warnings/Errors box and even select, copy (ctrl c) and paste (ctrl v) the contents if necessary.

**Related Topics:** *Input Data, Data Requirements, SaTScan Support.*

---

## Saving Analysis Parameters


Analysis parameters, specified on the Parameter tab dialog, can be saved and reused for future analyses. It is recommended that you save the parameters with a “.prm” file extension. The parameter file is stored in an ASCII text file format.

### To save analysis parameters

1. If the parameters have not previously been saved, select Save As from the File menu. A ‘Save Parameter File As’ dialog will open.
2. Select a directory location from the ‘Save In’ drop-down menu at the top of the dialog box.
3. Enter a name for your parameter file in the ‘File Name’ text box. It is recommended that the ‘Save As Type’ selection remain as Parameter Files (\*.prm).
4. Press the Save button.

Once the parameter file is initially saved, save changes to the file by selecting ‘Save’ on the File menu. The file will save without opening the ‘Save Parameter File As’ dialog.

### To open a saved parameter file

1. Select ‘Open’ from the File menu or click on the  button in the toolbar. A Select Parameter File dialog will open.
2. Locate the desired file using the Look in drop-down menu.
3. Once the file is located, highlight the file name by clicking on it.
4. Press the Open button.

A Parameter tab dialog will open containing the saved parameter settings. The location and name of the parameter file is listed in the title bar of this dialog.

**Related Topics:** *Specifying Analysis and Data Options, Basic SaTScan Features, Advanced Features, Batch Mode.*

---

## Parallel Processors

If you have parallel processors on your computer, SaTScan can take advantage of this by running different Monte Carlo simulations using different processors, thereby increasing the speed of the calculations. The default is that SaTScan will use all processors that the computer has. If you want to restrict the number, you can do that by clicking on Session > Execute Options, and selecting the maximum number of processors that SaTScan is allowed to use.

---

## Batch Mode

SaTScan is most easily run by clicking the Execute  button at the top of the SaTScan window, after filling out the various parameter fields in the Windows interface.

An alternative approach is to skip the windows interface and launch the SaTScan calculation engine directly by either:

1. Dragging a parameter file onto the 'SaTScanBatch.exe' executable.
2. Writing 'SaTScanBatch.exe \*.prm' in a batch file or at the command prompt, where \*.prm is the name of the parameter file.

By using the batch mode version, it is possible to write special software that incorporates the SaTScan calculation engine with other applications, such as an automated daily surveillance system for the early detection of disease outbreaks. To use SaTScan in this manner requires a reasonable amount of computer skills and sophistication.

When running SaTScan in batch mode, the parameter file may still be changed using the SaTScan windows interface. It is also possible to change the parameter manually using any text editor or automatically by using some other software product. If only a few parameter should change compared to what is in an existing parameter file, name that parameter file on the command prompt together with instruction on which parameters should change. The command line parameter values will then over-ride the parameter values specified in the parameter file.

When the batch mode version of SaTScan is run, the standard results file does not automatically pop up on the screen, but must be opened manually using any available text editor such as Notepad.

**Opportunity:** There are some parameter options that are not allowed when SaTScan is run under the windows interface but which can be set when run in batch mode. A few such examples are the number of Monte Carlo replications, the collection of ellipses used for the elliptic scan statistic, an unlimited number of multiple data set, the use of Gumbel approximated p-values for any probability model and the ability to write all the simulated data to a file (with the location IDs listed in alphabetical order). Parameter options not allowed by the windows interface have not all been thoroughly tested though, so there is some risk involved when running such analyses.

**Related Topics:** *Launching the Analysis, Basic SaTScan Features, Advanced Features, Saving Analysis Parameters.*

---

## Computing Time

The spatial and space-time scan statistics are computer intensive to calculate. The computing time depends on a wide variety of variables, and depending on the data set and the analytical options chosen; it could range from a few seconds to several days or weeks. The multinomial, ordinal and normal models are in general much more computer intensive than the other discrete scan statistics. Other than that, the three main things that increase the computing time is the number of locations in the coordinates and special grid files, the number of time intervals (for space-time analyses) and the number of data sets used.

### ***Single Data Set***

For a single data set, the computing time for one of the discrete scan statistics is approximately on the order of:

$$LGMT^k mS / P$$

where:

L = number of geographical data locations in the coordinates file (L=1 for purely temporal analyses)

G = number of geographical coordinates in the special grid file. If there is no such file, G=L.

M = maximum geographical cluster size, as a proportion of the population (  $0 < M = \frac{1}{2}$  , M=1 for a purely temporal analysis)

T = number of time intervals into which the temporal data is aggregated (T=1 for a purely spatial analysis)

m = maximum temporal cluster size, as a proportion of the study period (  $0 < m = 0.9$  , m=1 for purely spatial analysis)

S = number of Monte Carlo simulations

P = number of processors available on the computer for SaTScan use

k = 1 for purely spatial, prospective temporal and prospective space-time analyses without adjustments for earlier analyses

k = 2 for retrospective temporal and retrospective space-time analyses

The unit of the above formula depends on the probability model used and on the speed of the computer. When the total number of cases is very large compared to the number of locations and time intervals, the computing time for the discrete Poisson, Bernoulli and exponential models is instead on the order of:

$$CS / P$$

where:

C = the total number of cases

### ***Multiple Data Sets***

An analysis using multiple data sets is considerably more computer intensive than the analysis of a single data set. For the discrete Poisson, Bernoulli and exponential models, the computing time for two data sets

is much more than twice the time for a single data set. The computing time for  $D > 2$  data sets is approximately  $D/2$  times longer than the computing time for two data sets.

**Related Topics:** *Coordinates File, Grid File, Spatial Window Tab, Temporal Window Tab, Monte Carlo Replications, Early Termination of Simulations, Multiple Data Sets Tab.*

---

## Memory Requirements

SaTScan uses dynamic memory allocation. Depending on the nature of the input data, SaTScan will automatically choose one of two memory allocation schemes: the standard one and a special one for data sets with very many spatial locations but few time intervals and few simulations.

### Standard Memory Allocation

Using the standard memory allocation scheme, the amount of memory bytes needed for large data sets is approximately:

Discrete Poisson:	$ALGM + (12 + 4P) LTD + 8CRP$
Bernoulli:	$ALGM + (16 + 4P) LTD + 8CRP$
Space-Time Permutation:	$ALGM + (12 + 4P) LTD + 12CP + 8CRP$
Ordinal, Multinomial:	$ALGM + (4Y + 4YP + 4P) LTD$
Exponential:	$ALGM + (16 + 12P) LTD + 40IP + 8CRP$
Normal:	$ALGM + (20 + 16P) LTD + 32IP$
Normal, with weights:	$ALGM + (20 + 16P) LTD + 48IP$
Continuous Poisson:	$G + 16C + 24CP$

where

$L$  = the number of location IDs in the coordinates file ( $L=1$  for a purely temporal analysis)

$A = 2$  if  $L < 65,536$  and  $A = 4$  if  $L > 65,536$

$G$  = the number of coordinates in the grid file ( $G=L$  if no grid file is specified)

$M$  = maximum geographical cluster size, as a proportion of the population ( $0 < M = \frac{1}{2}$ ,  $M=1$  for a purely temporal analysis)

$T$  = number of time intervals into which the temporal data is aggregated ( $T=1$  for a purely spatial analysis)

$Y$  = the number of categories in the multinomial or ordinal model

$C$  = the total number of cases in the Poisson, Bernoulli, space-time permutation and exponential models

$I$  = the number of individual observations in the exponential and normal models

$R = 1$  when scanning for high rates only or low rates only,  $R=2$  when scanning for either high or low rates

$D$  = number of data sets

$P$  = number of processors available on the computer for SaTScan use

For purely spatial analyses and most space-time analyses, T is much less than G, as is D and P, so it is the SLGM expression to the left of the first plus sign above that is critical in terms of memory requirements for the discrete scan statistics. Table 2 provides estimates of the memory requirements when G=L, M=0.5 and T=1.

G=L	Memory Needed	G=L	Memory Needed
3,500	32Mb	44,000	2Gb
6,500	64Mb	63,000	4Gb
10,000	128Mb	89,000	16Gb
15,000	256Mb	126,000	32Gb
22,000	512Mb	178,000	64Gb
32,000	1Gb	250,000	128Gb

**Table 2: Approximate memory requirements for a purely spatial analysis when the maximum geographical cluster size is 50% of the population.**

### **Special Memory Allocation**

When the number of locations is very large while the number of cases, time intervals and simulations are not, SaTScan sometimes uses an alternative memory allocation scheme to reduce the total memory requirement. This selection is done automatically. The amount of memory needed for large data sets is then approximately the following number of bytes:

Discrete Poisson:	$(4S + 12 + 4P) LTD + 8CRS$
Bernoulli:	$(4S + 16 + 4P) LTD + 8CRS$
Space-Time Permutation:	$(4S + 12 + 4P) LTD + 12CP + 8CRS$
Ordinal, Multinomial:	$(4YS + 4Y + 4YP + 4P) LTD$
Exponential:	$(12S + 16 + 12P) LTD + 40IP + 8CRS$
Normal:	$(16S + 20 + 16P) LTD + 32IP$
Normal, with weights:	$(16S + 20 + 16P) LTD + 48IP$
Continuous Poisson:	not used

where S is the number of Monte Carlo simulations and the other variables are defined as above.

### **Insufficient Memory**

If there is insufficient memory available on the computer to run the analysis using either memory allocation scheme, there are several options available for working around the limitation:



- Close other applications.
- Aggregate the data into fewer data locations (reduce L).
- Decrease the number of circle centroids in the special grid file (reduce G).
- Reduce the upper limit on the circle size (reduce mg).
- Run the program on a computer with more memory.

It is highly desirable that there is sufficient RAM to cover all the memory needs, as SaTScan runs considerably slower when the swap file is used, so these techniques may also be used to avoid the swap file. Not all of these above options will work for all data sets. Please note that the following SaTScan options do not influence the demand on memory:

- The length of the study period.
- The maximum temporal cluster size.
- Type of space-time clusters to include in the analysis.

**Note:** The 32-bit windows operating system can allocate a maximum of 2 GBytes of memory to a single application, and that is hence the upper limit on the memory for the 32-bit windows version of SaTScan. The Linux version of SaTScan can be used to analyze larger data sets.

**Related Topics:** *Coordinates File, Grid File, Spatial Temporal and Space-Time Scan Statistics, Spatial Window Tab, Temporal Window Tab, Monte Carlo Replications, Multiple Data Sets Tab, Warnings and Errors.*

# Results of Analysis

As output, SaTScan creates one standard text based results file in ASCII format, two optional geographical format output files in KML or shapefile format and five optional columnformat output files in either ASCII or dBase format. Some of the optional files are useful when exporting output from SaTScan into other software such as a spreadsheet or a geographical information system.

All results file will be in the same directory and have the same name as the standard output file specified on the Output Tab, except for the extension.

**Related Topics:** *Output Tab, Spatial Output Tab, Standard Results File, Cluster Information File, Location Information File, Risk Estimates for Each Location, Simulated Log Likelihood Ratios, Analysis History File.*

---

## Standard Text-Based Results File (\*.out.\*)

The standard results file is automatically shown after the calculations are completed. It is fairly self-explanatory, but for proper interpretation it is recommended to read the section on statistical methodology, or even better, one of the methodological papers listed in the bibliography.

**SUMMARY OF DATA:** Use this to check that the input data files contain the correct number of cases, locations, etc.

**Total population** (discrete Poisson model): This is the average population during the study period.

**Annual rate per 100,000** (discrete Poisson model): This is calculated taking leap years into account and is based on the average length of a year of 365.2425. If calculated by hand ignoring leap years, the numbers will be slightly different, but not by much.

**Variance** (normal model): This is the variance for all observations in the data assuming a common mean.

**MOST LIKELY CLUSTER:** Summary information about the most likely cluster, that is, the cluster that is least likely to be due to chance.

**Radius:** When latitude and longitude are used, the radius of the circle is given in kilometers. When regular Cartesian coordinates are used, the radius of the circle is given in the same units as those used in the coordinates file.

**Population:** This is the average population in the geographical area of the cluster. The average is taken over the whole study period even when it is a space-time cluster whose temporal length is only a part of the study period.

**Relative Risk:** This is the estimated risk within the cluster divided by the estimated risk outside the cluster. It is calculated as the observed divided by the expected within the cluster divided by the observed divided by the expected outside the cluster. In mathematical notation, it is:

$$RR = \frac{c / E[c]}{(C - c) / (E[C] - E[c])} = \frac{c / E[c]}{(C - c) / (C - E[c])}$$

where  $c$  is the number of observed cases within the cluster and  $C$  is the total number of cases in the data set. Note that since the analysis is conditioned on the total number of cases observed,  $E[C]=C$ .

**Observed / Expected:** This is the observed number of cases within the cluster divided by the expected number of cases within the cluster when the null hypothesis is true, that is, when the risk is the same inside and outside the cluster. This means that it is the estimated risk within the cluster divided by the estimated risk for the study region as a whole. It is calculated as:  $c/E[c]$ .

For the continuous Poisson model, the expected count is an upper bound when the scanning window crosses the border of the spatial study region. That means that the Obs/Exp is a lower bound.

**Variance** (normal model): This is the estimated common variance for all observations in the, taking into account the different estimated means inside and outside the cluster. The weighted variance is adjusted for the weights when provided by the user.

**Time trend** (spatial variation in temporal trends): Provides the estimated time trends inside and outside the detected clusters on the log linear scale where the percent increase or percent decrease is constant over time.

**P-value:** The p-values are adjusted for the multiple testing stemming from the multitude of circles/cylinders corresponding to different spatial and/or temporal locations and sizes of potential clusters evaluated. This means that under the null-hypothesis of complete spatial randomness there is a 5% chance that the p-value for the most likely cluster will be smaller than 0.05 and a 95% chance that it will be bigger. Under the null hypothesis there will always be some area with a rate higher than expected just by chance alone. Hence, even though the most likely cluster always has an excess rate when scanning for areas with high rates, the p-value may actually be very close or identical to one.

**Recurrence Interval:** For prospective analyses, the recurrence interval<sup>20</sup> (or, null occurrence rate) is shown as an alternative to the p-value. The measure reflects how often a cluster of the observed or larger likelihood will be observed by chance, assuming that analyses are repeated on a regular basis with a periodicity equal to the specified time interval length. For example, if the observed p-value is used as the cut-off for a signal and if the recurrence interval is once in 14 months, then the expected number of false signals in any 14 month period is one.

If no adjustments are made for earlier analysis, then the recurrence interval is once in  $D/p$  days, where  $D$  is the number of days in each time interval. If adjustments are made for  $A-1$  earlier analyses, then the recurrence interval is once every  $D / [1 - (1-p)^{1/A}]$  days.

**SECONDARY CLUSTERS:** Summary information about other clusters detected in the data. The information provided is the same as for the most likely cluster. Only clusters with  $p < 1$  are displayed.

P-values listed for secondary clusters are calculated in the same way as for the most likely cluster, by comparing the log likelihood ratio of secondary clusters in the real data set with the log likelihood ratios of the most likely cluster in the simulated data sets. This means that if a secondary cluster is significant, it can reject the null hypothesis on its own strength without help of any other clusters. It also means that these p-values are conservative<sup>1</sup>.

**PARAMETER SETTINGS:** A reminder of the parameter settings used for the analysis.

**Additional results files:** The name and location of additional results files are provided, when applicable.

**Related Topics:** *Output Tab, Spatial Output Tab, Cluster Information File, Location Information File, Risk Estimates for Each Location, Simulated Log Likelihood Ratios, Cartesian Coordinates, Column Output Format.*

---

## Shapefile Geographical Output (\*.shp and \*.shx)

With the shapefile output, the detected SaTScan cluster locations can be directly exported to and shown in a wide variety of geographical information systems, including both free software such as Quantum GIS ([www.qgis.org](http://www.qgis.org)), TerraView ([www.dpi.inpe.br/terraview\\_eng/](http://www.dpi.inpe.br/terraview_eng/)) and R ([www.r-project.org](http://www.r-project.org)), as well as commercial products such as ArcGIS and SAS.

The shapefile output can only be used when the geographical coordinates are specified using latitudes and longitudes.

**Related Topics:** *Output Tab, Standard Results File, KML Output File, Spatial Output Tab, Cluster Information File, Geographical Coordinates File, Lat/Long Coordinates, Column Output Format.*

---

## Cluster Information File (\*.col.\*)

In the cluster information file, each cluster is on one line, with different information about the cluster in different columns. For each cluster there is information about the location and size of the cluster, its log likelihood ratio and the p-value. Except for the multinomial and ordinal models, and when multiple data sets are used, there is also information about the observed and expected number of cases, observed/expected and relative risk. For the multinomial and ordinal models, and for multiple data sets, these numbers depend on the data set and/or category, and the information is instead provided in the Stratified Cluster Information File.

The exact columns included in the file depend on the chosen analysis, as shown in Table 3, but is easily verified by comparison with the standard results file. The file will have the same name as the standard results file, but with the extensions \*.col.txt and \*.col.dbf respectively, and will be located in the same directory.

**Note:** While the standard results file only displays clusters with  $p < 1$ , this file will also display clusters with  $p = 1$ . Note that these p-values are adjusted for multiple testing, so even if  $p = 1$ , the cluster may still have a fairly high relative risk.

**Related Topics:** *Cluster Cases Information File, Location Information File, Output Tab, Results of Analysis, Standard Results File*

Output Variable	dBase Name	Discrete Poisson, Bernoulli, Circular, Lat/Long, One Data Set	Discrete Poisson, Bernoulli, Circular, Cartesian, One Data Set	Discrete Poisson, Bernoulli, Circular, Cartesian 5 Dimensions, One Data Set	Space-Time Permutation, Circular, Lat/Long, One Data Set	Exponential, Circular, Lat/Long, One Data Set	Discrete Poisson, Bernoulli, Elliptic, Cartesian, One Data Set	Ordinal, Multinomial, Circular, Lat/Long, One Data Set	Discrete Poisson, Bernoulli, Circular, Lat/Long, Multiple Data Sets	Normal, Circular, Lat/Long, One Data Set	Normal, Weighted, Circular, Lat/Long, One Data Set	Continuous Poisson
Cluster Number	CLUSTER	1	1	1	1	1	1	1	1	1	1	1
Central Location ID	LOCATION_ID	2	2	2	2	2	2	2	2	2	2	2
Latitude	LATITUDE	3	-	-	3	3	-	3	3	3	3	-
Longitude	LONGITUDE	4	-	-	4	4	-	4	4	4	4	-
X-coordinate	X	-	3	3	-	-	3	-	-	-	-	3
Y-coordinate	Y	-	4	4	-	-	4	-	-	-	-	4
Z1-coordinate	Z1	-	-	5	-	-	-	-	-	-	-	-
Zn-coordinate	Zn	-	-	6,7	-	-	-	-	-	-	-	-
Circle Radius	RADIUS	5	5	8	5	5	-	5	5	5	5	5
Ellipse, Length of Minor Axis	E_MINOR	-	-	-	-	-	5	-	-	-	-	-
Ellipse, Length of Major Axis	E_MAJOR	-	-	-	-	-	6	-	-	-	-	-
Ellipse Angle	E_ANGLE	-	-	-	-	-	7	-	-	-	-	-
Ellipse Shape	E_SHAPE	-	-	-	-	-	8	-	-	-	-	-
Cluster Start Date	START_DATE	6	6	9	6	6	9	6	6	6	6	-
Cluster End Date	END_DATE	7	7	10	7	7	10	7	7	7	7	-
# Location IDs	NUMBER_LOC	8	8	11	8	8	11	8	8	8	8	6
Log Likelihood Ratio	LLR	9	9	12	-	9	12	9	9	9	9	7
Test Statistic	TEST_STAT	-	-	-	9	-	13	-	-	-	-	-
P-Value of Cluster	P_VALUE	10	10	13	10	10	14	10	10	10	10	8
Observed Cases	OBSERVED	11	11	14	11	11	15	-	-	11	11	9
Expected Cases	EXPECTED	12	12	15	12	12	16	-	-	-	-	10
Observed / Expected	ODE	13	13	16	13	13	17	-	-	-	-	11
Relative Risk	REL_RISK	14	14	17	-	-	18	-	-	-	-	-
Total Weights	WEIGHT_IN	-	-	-	-	-	-	-	-	-	12	-
Mean Inside	MEAN_IN	-	-	-	-	-	-	-	-	12	13	-
Mean Outside	MEAN_OUT	-	-	-	-	-	-	-	-	13	14	-
Variance	VARIANCE	-	-	-	-	-	-	-	-	14	15	-
Standard deviation	STD	-	-	-	-	-	-	-	-	15	16	-
Weighted Mean Inside	W_MEAN_IN	-	-	-	-	-	-	-	-	-	17	-
Weighted Mean Outside	W_MEAN_OUT	-	-	-	-	-	-	-	-	-	18	-
Weighted Variance	W_VARIANCE	-	-	-	-	-	-	-	-	-	19	-
Weighted Standard deviation	W_STD	-	-	-	-	-	-	-	-	-	20	-

*Table 3: Content of the cluster information output file, with dBase variable names and examples of column ordering for a few different types of analyses.*

---

## Stratified Cluster Information File (\*.sci.\*)

In the stratified cluster information file, there is one line for each ordinal/multinomial category, in each data set, for each cluster. For each cluster/category/data set combination, there is one column each for the observed number of cases, the expected number of cases, observed divided by expected and sometimes, the relative risk. If neither the multinomial model, ordinal model nor multiple data sets are used, then there is only one line for each cluster, and there is no information in this file that is not provided in the Cluster Information File.

File format: <Cluster #><Data Set #><Category #><Observed><Expected><Obs/Exp><RR>

The file will have the same name as the standard results file, but with the extensions \*.sci.txt and \*.sci.dbf respectively, and will be located in the same directory.

**Related Topics:** *Cluster Information File, Location Information File, Output Tab, Results of Analysis, Standard Results File.*

---

## Location Information File (\*.gis.\*)

As an option, a special output file may be created describing the various clusters in a way that is easy to incorporate into a geographical information system (GIS). This file may be requested in ASCII and/or dBase format, and can be accessed using any text editor or spreadsheet program. It will have the same name as the results file, but with the extensions \*.gis.txt and \*.gis.dbf respectively, and it will be located in the same directory. This file has one row for each location belonging to a cluster. The columns shown depend on the chosen analysis, including among other the following information:

- <Location ID>
- <Cluster Number>
- <P-Value of Cluster>
- <Observed Cases in Cluster>
- <Expected Cases in Cluster>
- <Observed/Expected in Cluster>
- <Observed Cases in Location>
- <Expected Cases in Location>
- <Observed/Expected in Location>

**Note:** The second, third, fourth, fifth and sixth column entries are the same for all locations belonging to the same cluster.

**Related Topics:** *Output Tab, Results of Analysis, Standard Results File, Cluster Information File.*

---

## Risk Estimates for Each Location File (\*.rr.\*)

If the option to include risk estimates for each location is selected, a file with a list of all data locations and the corresponding number of observed cases, number of expected cases, the observed/expected ratio and the relative risk for each location is provided. This may be useful when examining a cluster area in more detail. The information is purely descriptive. There is one line for each Location ID, and the content of the five columns is as follows:

**<Location ID><Observed Cases> <Expected Cases><Observed/Expected><Relative Risk>**

This file may be accessed using any text editor or spreadsheet program. It will have the same name as the results file, but with the extension \*.rr.txt or \*.rr.dbf, and it will be located in the same directory. The file is only available for the discrete scan statistic, and hence, not for the continuous Poisson model.

**Related Topics:** *Output Tab, Results of Analysis, Standard Results File.*

---

## Simulated Log Likelihood Ratios File (\*.llr.\*)


The log likelihood ratio test statistics from the random data sets are not provided as part of the standard output. If desired, they can be printed to a special file which by default has the same name as the output file but with the extension \*.llr.txt or \*.llr.dbf. There is typically no need for this file, but it can be useful for statistical researchers who may be interested in the distributional properties of the scan statistic under various scenarios.

**Related Topics:** *Output Tab, Results of Analysis, Standard Results File, Monte Carlo Replications.*

# Miscellaneous

---

## New Versions

To check whether there is a later version than the one you are currently using, simply click on the update button  on the tool bar. If a newer version exists, you will be asked whether you want to automatically download and install it. At any given time, it is also possible to download the latest version of SaTScan from the World Wide Web at '<http://www.satscan.org/>'.

**Related Topics:** *Download and Installation.*

---

## Analysis History File

In the analysis history file, SaTScan automatically maintains a log of all the SaTScan analyses conducted. Included in the log is an assigned analysis number together with information about the time of the analysis, parameter settings, a very brief summary of the results, as well as the name of the standard results file created.

The analysis history is in a dBase file with the name AnalysisHistory.dbf, located in the same directory as the SaTScan executable. It can be opened and read using most database and spreadsheet software, including Excel. You can erase the file at any time. A new file will then be created the next time you run SaTScan, starting the list of analyses from scratch.

**Related Topics:** *Running SaTScan, Results of Analysis.*

---

## Random Number Generator

The choice of random number generator is critical for any software creating simulated data. SaTScan uses a Lehmer random number generator<sup>26</sup> with modulus  $2^{31}-1 = 2147483647$  and multiplier 48271, which is known to perform well<sup>27</sup>.

**Related Topics:** *Monte Carlo Replications.*



---

## Contact Us

Please direct technical questions about installation and running the program, as well as the web site, to:

[techsupport@satscan.org](mailto:techsupport@satscan.org)

Please direct substantive questions about the statistical methods and suggestions about new features to:

Martin Kulldorff, Associate Professor, Biostatistician  
Department of Population Medicine  
Harvard Medical School and Harvard Pilgrim Health Care Institute  
133 Brookline Avenue, 6th Floor, Boston, MA 02215, USA  
Email: [kulldorff@satscan.org](mailto:kulldorff@satscan.org)

---

## Acknowledgements

### Financial Support

**National Cancer Institute**, Division of Cancer Prevention, Biometry Branch [SaTScan v1.0, 2.0, 2.1]

**National Cancer Institute**, Division of Cancer Control and Population Sciences, Statistical Research and Applications Branch [SaTScan v3.0 (part), 6.1 (part), 9.2]

**Alfred P. Sloan Foundation**, through a grant to the New York Academy of Medicine (Farzad Mostashari, PI) [SaTScan v3.0 (part), 3.1, 4.0, 5.0, 5.1]

**Centers for Disease Control and Prevention**, through Association of American Medical Colleges Cooperative Agreement award number MM-0870 [SaTScan v6.0, 6.1 (part)]

**National Institute of Child Health and Development**, through grant #R01HD048852 [7.0,8.0,v9.0(part)]

**National Institute of General Medical Sciences**, through Modeling Infectious Disease Agent Study (MIDAS) grant #U01GM076672 [v9.0 (part),9.1]

Their financial support is greatly appreciated. The contents of SaTScan are the responsibility of the developer and do not necessarily reflect the official views of funders.

### Comments and Suggestions

Feedback from users is greatly appreciated. Very valuable suggestions concerning the SaTScan software have been received from many individuals, including:

Allyson Abrams, Harvard Medical School & Harvard Pilgrim Health Care

Frank Boscoe, New York State Health Department

Eric Feuer, National Cancer Institute

Laurence Freedman, National Cancer Institute

David Gregorio, University of Connecticut

Göran Gustafsson, Karolinska Institute, Sweden

Jessica Hartman, New York Academy of Medicine

Richard Heffernan, New York City Department of Health

Kevin Henry, New Jersey Department of Health  
Ulf Hjalmar, Östersund Hospital, Sweden  
Richard Hoskins, Washington State Department of Health  
Lan Huang, National Cancer Institute  
Ahmedin Jemal, American Cancer Society  
Inkyung Jung, Harvard Medical School & Harvard Pilgrim Health Care  
Ann Klassen, Johns Hopkins University  
Ken Kleinman, Harvard Medical School & Harvard Pilgrim Health Care  
Sanjaya Kumar, New York State Health Department  
Kristina Metzger, New York City Department of Health  
Barry Miller, National Cancer Institute  
Farzad Mostashari, New York City Department of Health  
Lloyd Mueller, Connecticut Tumor Registry  
Karen Olson, Children's Hospital, Boston  
Linda Pickle, National Cancer Institute  
Simon Read, University of Sheffield  
Tom Richards, Centers for Disease Control and Prevention  
Gerhard Rushton, University of Iowa  
Joseph Sheehan, University of Connecticut  
Tom Talbot, New York State Health Department  
Toshiro Tango, National Institute of Public Health, Japan  
Jean-François Viel, Université de Franche-Comté, France  
Shihua Wen, University of Maryland

# Frequently Asked Questions

---

## Input Data

- 1. I tried running SaTScan using one of the sample data sets, and all went well, but when I try it on my own data there is an error. What should I do?**

SaTScan makes sure that the input data is compatible with each other, and with the options specified on the windows interface. For example, it complains if there is a location ID in the case file that is not present in the coordinates file, as it must know where to localize those cases. For most data sets there is some need for data cleaning and SaTScan is designed to help with this process by spotting and pointing out any inconsistencies found.

- 2. I have constructed the ASCII input files exactly according to the description in the SaTScan User Guide, but SaTScan complains that they are not in the correct format. What is wrong?**

The most likely explanation is that the files are in UNICODE rather than ASCII format. Just convert to ASCII and it should work.

- 3. In my data, there is zero or only one case in most locations. Can I use SaTScan for such sparse data?**

Yes, you certainly can. One of the main reasons for using SaTScan is to avoid arbitrary geographical aggregation of the data, letting the scan statistic consider different smaller or larger aggregations through its continuously moving window. With finer geographical resolution of the input data, SaTScan can evaluate more different cluster locations and sizes without restrictions imposed by administrative geographical boundaries, minimizing assumptions about the geographical cluster location and size.

- 4. If my data is sparse, won't the rates be statistically unstable?**

The stability of rates does not depend on the geographical resolution of the input data, but on the population size of the circles constructed by SaTScan.

- 5. What is the minimum number of spatial locations needed to run SaTScan?**

The purely temporal scan statistic can be run with only one geographical location. The space-time scan statistic needs at least two locations. With only two locations, the space-time scan statistic will look for temporal clusters in either or both of the locations. Technically, the purely spatial scan statistic can also be run using only two geographical locations, providing correct inference. There is no point using a purely spatial scan statistic for such data though, for which a regular chi-square statistic can be used instead, as there is no multiple testing to adjust for. With three locations or more, the fundamental scan statistic concept of including different combinations of locations into the potential clusters is being utilized. In most practical applications though, the spatial and space-time scan statistics are used for data sets with

hundreds or thousands of geographical locations. If there is a choice, less spatial aggregation of the data is typically better, which means more geographical locations.

---

## Analysis

### **6. With latitude/longitude coordinates, what planar projection is used?**

No projection is used. SaTScan draws perfect circles on the spherical surface of the earth.

### **7. When should I use the Bernoulli versus the Poisson model?**

Use the Bernoulli model when you have binary data, such as cases and controls, late and early stage cancer or people with and without a disease. Use the Poisson model when you have cases and a background population at risk, such as population numbers from the census.

### **8. SaTScan adjusts for categorical covariates, but I want to adjust for a continuous variable. Is that possible?**

One way to do this is to categorize the continuous variable. A better approach is to (i) calculate the adjustment using a regular statistical software package such as SAS, (ii) use the result from that analysis to calculate the covariate adjusted expected number of cases at each location, and (iii) use these expected values instead of the population in the population file. With this approach, there should not be any covariates in either the case or the population files.

### **9. What should I use as the maximum geographical cluster size? Is that an arbitrary choice?**

If you don't want the choice to be arbitrary, choose 50% of the population as the maximum geographical cluster size. SaTScan will then evaluate very small and very large clusters, and everything in-between. To find a good collection of non-overlapping cluster, use the Gini index feature.

### **10. Why can't I select a maximum geographical cluster size that is larger than 50% of the population?**

Clusters of excess risk that are larger than 50% of the population at risk are better viewed as cluster with lower risk outside the scanning window, and the area outside will always have a very irregular geographical shape. If there is interest in clusters with lower risk than expected, it is more appropriate to select the low rates option on the analysis tab.

### **11. I have memory problems when running SaTScan. What should I do?**

Make sure you are running SaTScan in 64-bit mode. For this you must (i) have a 64 bit computer, (ii) run SaTScan version 8 or later, and (iii) have 64-bit Java installed on your computer.

---

## Results

### **12. I get an error stating that the output file could not be created. Why?**

In Windows, permission to write to the "Program Files" folder is given only to administrators and power users of that machine. If the output file path includes the "Program Files" folder and you do not have administrative or power user privileges on your computer, Windows prevents SaTScan from creating the output file in the designated location. The solution is to specify a different output file name using a different directory.

### **13. Since the SaTScan results are based on Monte Carlo simulated random data, why are the p-values the same when I run the analysis twice?**

All computer-based simulations are based on pseudo-random number generators. When the same seed is used, exactly the same sequence of pseudo-random numbers will be generated. Since SaTScan uses the same seed for every run, you obtain the same result for two runs when the input data is the same.

### **14. I ran exactly the same data using two different versions of SaTScan, but the p-values are different. Why? Which one is the correct one?**

Compared to v2.1, the pseudo-random number generation is done slightly differently in SaTScan v3.0 and later, typically resulting in slightly different p-values. In earlier version, SaTScan defined overlapping clusters based on whether the two circles were overlapping. In SaTScan v5.0 and later, two clusters overlap if they have at least one location ID in common. These two definitions are usually the same, but in rare cases they may be different. If you were running the Poisson model, another possible reason for the difference is that SaTScan v5.0 and later uses a more precise algorithm for calculating the expected number of cases when the population dates in the population file are specified using days rather than months or years.

While p-values from all versions are valid and correct, only one p-value should be used. We recommend always using the p-value that was calculated first.

---

## Interpretation

### **15. In SaTScan, after adjusting for population density and covariates such as age, the null-hypothesis is complete spatial randomness. For most disease data that is not true. Does this mean that the null hypothesis is wrong?**

When accepting the notion of statistical hypothesis testing one must also accept the fact that the null hypothesis is never true. For example, when comparing the efficacy of two different surgical procedures in a clinical trial we know for sure that their efficacy cannot be equal, but we still use equality as the null hypothesis since we are interested in finding out whether one is better than the other. Likewise, with geographical data we know that disease risk is not the same everywhere but we still use it as the null hypothesis since we are interested in finding locations with excess risk. Hence, the null hypothesis is wrong in the sense that we know it is not true but it is not wrong in the sense that we should not use it.

- 16. Does SaTScan assume that there is no spatial auto-correlation in the data? (Note: Spatial auto-correlation means that the location of disease cases is dependent on the location of other disease cases, such as with an infectious disease where an infected individual is likely to infect those living close by.)**

No, SaTScan does not assume that there is no spatial auto-correlation in the data. Rather, it is a test of whether there is spatial auto-correlation or other divergences from the null hypothesis. In this sense it is equivalent to a statistical test for normality, which does not assume that the data is normally distributed but tests whether it is.

- 17. If I am interested in whether there is spatial auto-correlation in the data, why should I use the spatial scan statistic rather than a traditional spatial auto-correlation test?**

If you are only interested in whether there is spatial auto-correlation or not, but don't care about cluster locations, there are tests for spatial auto-correlation / global clustering that have higher power than the spatial scan statistic and should be used instead. The spatial scan statistic should be used when you are interested in the detection and statistical significance of local clusters.

- 18. In spatial statistics, is it not always important to adjust for spatial auto-correlation? This cannot be done in SaTScan.**

Whether to adjust for spatial auto-correlation depends on the question being asked from the data. As an example, let's assume that we have geographical data on people who get sick due to food poisoning. In such data there is clearly spatial auto-correlation, since bad food sold at restaurants or grocery stores are often sold to multiple customers, many of who will live in the same neighborhood.

If we are doing spatial regression trying to determine what neighborhood characteristics such as mean income, house values, educational levels or ethnic origin contribute to a higher risk for food poisoning, it is critical to adjust for the spatial auto-correlation in the data. If not, the confidence in the risk relationships will be overestimated with biased p-values that are too small, providing 'statistically significant' results when none exist. Here, the null hypothesis should be that there is spatial auto-correlation and the alternative hypothesis that there are geographical differences in the risk of food poisoning.

On the other hand, if we are interested in quickly detecting food poisoning outbreaks, we should not adjust for the spatial auto-correlation since we are interested in detecting clusters due to such correlation, and if they are adjusted away, important clusters may go undetected. Here, the null hypothesis is that the food poisoning cases are geographically randomly distributed (adjusted for population density etc.) and the alternative hypothesis is that there is some clustering either due to differences in underlying risk factors or spatial auto-correlation. Once the location of a cluster has been detected, it is for the local health officials to determine the source of the cluster to prevent further illness.

- 19. If there are multiple clusters in the data, does that mean that the p-values are more likely to be significant than their 0.05 nominal significance level suggests, so that chance clusters are detected too often?**

No. The opposite is actually true. Looking at United States mortality, suppose we have 1000 cases of a disease in Seattle and 30 in New York City. Seattle is clearly a significant cluster but 30 cases in New York City out of 1030 in all of the USA is not exceptional since the City has about 3 percent of the U.S. population. If we accept that there is a cluster in Seattle though, and if we adjust for that by removing Seattle from the analysis, then 30 cases in the City out of 30 nationwide is statistically significant. This is similar to a regular multiple regression, where if we adjust for one variable, another variable may suddenly become statistically significant. Note that the opposite is also true. If we remove an area with significantly fewer cases than expected, then a significant cluster with an excess number of cases may become non-significant.

- 20. For count data, the spatial scan statistic uses a particular alternative hypothesis with an excess risk in a circular cluster, where the number of cases follows a Poisson or Bernoulli distribution. Does this mean that it can only be used to detect such alternative hypotheses?**

Many proposed and widely used test statistics do not specify an alternative hypothesis at all. This neither means that they cannot be used for any alternative hypotheses nor that they are good for all alternatives. Likewise, if an explicit alternative is defined, as with the spatial scan statistic, that does not mean that it cannot be used for other alternative hypotheses as well. It is simply a question of the test statistic having good power for some alternative hypotheses and low power for other. The advantage of having a well-specified alternative is that it gives some information about the alternatives for which the test can be expected to have good power.

- 21. For the exponential (normal) model, it is assumed that the survival times follow an exponential (normal) distribution. Are the results biased if the survival times follow a different distribution?**

No matter which distribution generated the survival times, the p-values from the statistical inference are still valid and unbiased. This is because rather than generating the random data from an exponential distribution, each random data is a spatial permutation of the survival times. A greatly misspecified distribution may lead to a loss in power though. For example, if the data is Bernoulli distributed, the exponential model has less power to detect a cluster than the Bernoulli model. For continuous distributions such as gamma and lognormal, the exponential model has been shown to work well. The same reasoning is true with respect to the normal model.

---

## Operating Systems

- 22. Is SaTScan available for Windows/Mac/Linux?**

The SaTScan software for Windows, Mac and Linux can be downloaded from the [www.satscan.org](http://www.satscan.org) web site.

# SaTScan Bibliography

Different SaTScan analysis options were developed at different times and they are described in different scientific publications. The following bibliography contains selected papers and reports intended to help you find information on the following:

1. Find the methodological paper(s) in which the various analysis options are presented and discussed in more detail than what is available here in the SaTScan User Guide.
2. Find applications in different scientific areas.
3. Determine the relevant scientific papers to cite.

---

## Suggested Citations

The SaTScan software may be used freely, with the requirement that proper references are provided to the scientific papers describing the statistical methods. For the most common analyses, the suggested citations are:

*Bernoulli, Discrete Poisson and Continuous Poisson Models:* Kulldorff M. A spatial scan statistic. Communications in Statistics: Theory and Methods, 26:1481-1496, 1997. [[online](#)]

*Space-Time Permutation Model:* Kulldorff M, Heffernan R, Hartman J, Assunção RM, Mostashari F. A space-time permutation scan statistic for the early detection of disease outbreaks. PLoS Medicine, 2:216-224, 2005. [[online](#)]

*Multinomial Model:* Jung I, Kulldorff M, Richard OJ. A spatial scan statistic for multinomial data. Statistics in Medicine, 2010, epub. [[online](#)]

*Ordinal Model:* Jung I, Kulldorff M, Klassen A. A spatial scan statistic for ordinal data. Statistics in Medicine, 2007; 26:1594-1607. [[online](#)]

*Exponential Model:* Huang L, Kulldorff M, Gregorio D. A spatial scan statistic for survival data. Biometrics, 2007; 63:109-118. [[online](#)]

*Normal Model without Weights:* Kulldorff M, Huang L, Konty K. A scan statistic for continuous data based on the normal probability model. International Journal of Health Geographics, 2009, 8:58. [[online](#)]

*Normal Model with Weights:* Huang L, Huang L, Tiwari R, Zuo J, Kulldorff M, Feuer E. Weighted normal spatial scan statistic for heterogeneous population data. Journal of the American Statistical Association, 2009, 104:886-898. [[online](#)]

*Software:* Kulldorff M. and Information Management Services, Inc. SaTScan™ v8.0: Software for the spatial and space-time scan statistics. <http://www.satscan.org/>, 2009.

Users of SaTScan should in any reference to the software note that: “SaTScan™ is a trademark of Martin Kulldorff. The SaTScan™ software was developed under the joint auspices of (i) Martin Kulldorff, (ii) the National Cancer Institute, and (iii) Farzad Mostashari of the New York City Department of Health and Mental Hygiene.”

**Related Topics:** *SaTScan Bibliography, Methodological Papers.*



---

# SaTScan Methodology Papers

## **Statistical Methodology**

### *General Statistical Theory, Bernoulli and Poisson Models*

1. Kulldorff M. A spatial scan statistic. Communications in Statistics: Theory and Methods, 1997; 26:1481-1496. [\[online\]](#)

### *Spatial Scan Statistic, Bernoulli Model*

2. Kulldorff M, Nagarwalla N. Spatial disease clusters: Detection and inference. Statistics in Medicine, 1995; 14:799-810. [\[online\]](#)

### *Retrospective Space-Time Scan Statistic*

3. Kulldorff M, Athas W, Feuer E, Miller B, Key C. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. American Journal of Public Health, 1998; 88:1377-1380. [\[online\]](#)

### *Prospective Space-Time Scan Statistic*

4. Kulldorff M. Prospective time-periodic geographical disease surveillance using a scan statistic. Journal of the Royal Statistical Society, 2001; A164:61-72. [\[online\]](#)

### *Space-Time Permutation Model*

5. Kulldorff M, Heffernan R, Hartman J, Assunção RM, Mostashari F. A space-time permutation scan statistic for the early detection of disease outbreaks. PLoS Medicine, 2005; 2:216-224. [\[online\]](#)

### *Multinomial Model*

6. Jung I, Kulldorff M, Richard OJ. A spatial scan statistic for multinomial data. Statistics in Medicine, 2010, epub. [\[online\]](#)

### *Ordinal Model*

7. Jung I, Kulldorff M, Klassen A. A spatial scan statistic for ordinal data. Statistics in Medicine, 2007; 26:1594–1607. [\[online\]](#)

### *Exponential Model*

8. Huang L, Kulldorff M, Gregorio D. A spatial scan statistic for survival data. Biometrics, 2007, 63:109-118. [\[online\]](#)

### *Normal Model*

9. Kulldorff M, Huang L, Konty K. A scan statistic for continuous data based on the normal probability model. International Journal of Health Geographics, 2009, 8:58. [\[online\]](#)

### *Weighted Normal Model*

10. Huang L, Huang L, Tiwari R, Zuo J, Kulldorff M, Feuer E. Weighted normal spatial scan statistic for heterogeneous population data. *Journal of the American Statistical Association*, 2009, 104:886-898. [[online](#)]

### *Spatial Variation in Temporal Trends*

11. Manuscript in preparation.

### *Multivariate Scan Statistic*

12. Kulldorff M, Mostashari F, Duczmal L, Yih K, Kleinman K, Platt R. Multivariate spatial scan statistics for disease surveillance. *Statistics in Medicine*, 2007, 26:1824-1833. [[online](#)]

### *Elliptic Scanning Window*

13. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. *Statistics in Medicine*, 2006, 25:3929-3943. [[online](#)]

### *Isotonic Spatial Scan Statistic*

14. Kulldorff M. An isotonic spatial scan statistic for geographical disease surveillance. *Journal of the National Institute of Public Health*, 1999;48:94-101. [[online](#)]

### *Monte Carlo Hypothesis Testing*

15. Dwass M. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 1957; 28:181-187
16. Besag J, Clifford J. Sequential Monte Carlo p-values. *Biometrika*, 1991; 78:301-330.
17. Silva I, Assunção RM, Costa M. Power of the sequential Monte Carlo test. *Sequential Analysis*, 2009; 28:163-174.

### *Gumbel P-Values*

18. Abrams A, Kleinman K, Kulldorff M. Gumbel based p-value approximations for spatial scan statistics.. *International Journal of Health Geographics* 2010, 9:61. [[online](#)]
19. Read S, Bath PA, Willett P, Maheswaran R. A study on the use of Gumbel approximation with the Bernoulli spatial scan statistic. *Statistics in Medicine*, 2013.

### *Recurrence Intervals*

20. Kleinman K, Lazarus R, Platt R. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *American Journal of Epidemiology*, 159:217-24, 2004.

## **Adjustments**

### *Adjusting for Covariates*

References [1] and [8] above, plus:

21. Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast cancer in northeastern United States: A geographical analysis. *American Journal of Epidemiology*, 146:161-170, 1997. [[online](#)]
22. Kleinman K, Abrams A, Kulldorff M, Platt R. A model-adjusted space-time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection*, 2005, 133:409-419.
23. Klassen A, Kulldorff M, Curriero F. Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *International Journal of Health Geographics*, 2005, 4:1. [[online](#)]

### *Iterative Scan Statistics, Adjusting for More Likely Clusters*

24. Zhang Z, Kulldorff M, Assunção R. Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics*, 2010, 642379.

## **Computational Aspects**

### *Algorithms*

25. Kulldorff M. Spatial scan statistics: Models, calculations and applications. In Balakrishnan and Glaz (eds), *Recent Advances on Scan Statistics and Applications*. Boston, USA: Birkhäuser, 1999. [[online](#)]

### *Random Number Generator*

26. Lehmer DH. Mathematical methods in large-scale computing units. In *Proceedings of the second symposium on large scale digital computing machinery*. Cambridge, USA: Harvard Univ. Press, 1951.
27. Park SK, Miller KW. Random number generators: Good ones are hard to find. *Communications of the ACM*, 31:1192-1201, 1988.

### *Macros*

28. Abrams AM, Kleinman KP. A SaTScan (TM) macro accessory for cartography (SMAC) package implemented with SAS (R) software. *International Journal of Health Geographics*, 6:6,2007. [[online](#)]

## **Reporting, Visualization and Mapping**

29. Boscoe FP, McLaughlin C, Schymura MJ, Kielb CL. Visualization of the spatial scan statistic using nested circles. *Health and Place*, 9:273-277, 2003.
30. North American Association of Central Cancer Registries, SaTScan-Google Earth Cluster Viewer, [[online](#)]
31. Han J, et al. Determining optional cluster reporting sizes for spatial scan statistics. Manuscript in preparation, 2013.

## **Methods Evaluations and Comparisons**

32. Kulldorff M, Tango T, Park P. Power comparisons for disease clustering tests. *Computational Statistics and Data Analysis*, 42:665-684, 2003.
33. Song C, Kulldorff M. Power evaluation of disease clustering tests. *International Journal of Health Geographics*, 2:9, 2003. [[online](#)]
34. Kulldorff M, Zhang Z, Hartman J, Heffernan R, Huang L, Mostashari F. Evaluating disease outbreak detection methods: Benchmark data and power calculations. *Morbidity and Mortality Weekly Report*, 53:144-151, 2004. [[online](#)]
35. Nordin J, Goodman M, Kulldorff M, Ritzwoller D, Abrams A, Kleinman K, Levitt MJ, Donahue J, Platt R. Using modeled anthrax attacks on the Mall of America to assess sensitivity of syndromic surveillance. *Emerging Infectious Diseases*, 11:1394-1398, 2005. [[online](#)]
36. Ozdenerol E, Williams BL, Kang SY, Magsumbol MS. Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters. *International Journal of Health Geographics*, 4:19, 2005. [[online](#)]
37. Costa MA, Assunção RM. A fair comparison between the spatial scan and Besag-Newell disease clustering tests. *Environmental and Ecological Statistics*, 12:301-319, 2005.
38. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4:11, 2005. [[online](#)]
39. Kulldorff M, Song C, Gregorio D, Samociuk H, DeChello L. Cancer map patterns: Are they random or not? *American Journal of Preventive Medicine*, 30:S37-49, 2006. [[online](#)]
40. Duczmal L, Kulldorff M, Huang L. Evaluation of spatial scan statistics for irregular shaped clusters. *Journal of Computational and Graphical Statistics*, 15:428-442, 2006.
41. Aamodt G, Samuelsen SO, Skrondal A. A simulation study of three methods for detecting disease clusters. *International Journal of Health Geographics*, 5:15, 2006. [[online](#)]
42. Jackson MC, Huang L, Luo J, Hachey M, Feuer E. Comparison of tests for spatial heterogeneity on data with global clustering patterns and outliers. *International Journal of Health Geographics*. 2009;8:55. [[online](#)]
43. Wheeler DC. A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996–2003. *International Journal of Health Geographics*. 2007;6:13. [[online](#)]
44. Goujon-Bellec S, Demoury C, Guyot-Goubin A, Hémon D, Clavel J. Detection of clusters of a rare disease over a large territory: performance of cluster detection methods. *International Journal of Health Geographics*. 2011;10:53. [[online](#)]

**Related Topics:** *SaTScan Bibliography, Selected Applications by Field of Study, Suggested Citation.*

---

## Selected SaTScan Applications by Field of Study

### ***Respiratory Infectious Diseases***

45. Andrade AL, Silva SA, Martelli CM, Oliveira RM, Morais Neto OL, Siqueira Junior JB, Melo LK, Di Fabio JL. Population-based surveillance of pediatric pneumonia: use of spatial analysis in an urban area of Central Brazil. *Cadernos de Saúde Pública*. 20: 411-421, 2004. [[online](#)]
46. Bakker MI, Hatta M, Kwenang A, Faber WR, van Beers SM, Klatser PR, Oskam L. Population survey to determine risk factors for *Mycobacterium leprae* transmission and infection. *International Journal of Epidemiology*, 33: 1329-1336, 2004.
47. Elias J, Harmsen D, Claus H, Hellenbrand W, Frosch M, Vogel U. Spatiotemporal analysis of invasive meningococcal disease, Germany. *Emerging Infectious Diseases*, 12:1689-1695, 2006. [[online](#)]
48. Oeltmann JE, Varma JK, Ortega L, Liu Y, O'Rourke T, Cano M, Harrington T, Toney S, Jones W, Karuchit S, Diem L, Rienthong D, Tappero JW, Ijaz K, Maloney S. Multidrug-Resistant Tuberculosis Outbreak among US-bound Hmong Refugees, Thailand, 2005. *Emerging Infectious Diseases*, 14:1715-1721, 2008. [[online](#)]
49. Fischer EAJ, Pahan D, Chowdhury SK, Oskam L, Richardus JH. The spatial distribution of leprosy in four villages in Bangladesh: An observational study. *BMC Infectious Diseases*, 8:125, 2008.
50. Liang L, Xu B, Chen Y, Liu Y, Cao W, Fang L, Feng L, Goodchild MF, Gong P. Combining spatial-temporal and phylogenetic analysis approaches for improved understanding on global H5N1 transmission. *PLoS One*. 5:e13575, 2010. [[online](#)]

### ***Food and Water Borne Diseases***

51. Cruz Payão Pellegrini D. Análise espaço-temporal da leptospirose no município do Rio de Janeiro (1995-1999). Rio de Janeiro: Fundação Oswaldo Cruz, 2002. [[online](#)]
52. Enemark HL, Ahrens P, Juel CD, Petersen E, Petersen RF, Andersen JS, Lind P, Thamsborg SM. Molecular characterization of Danish *Cryptosporidium parvum* isolates. *Parasitology*, 125:331-341, 2002.
53. Sauders BD, Fortes ED, Morse DL, Dumas N, Kiehlbauch JA, Schukken Y, Hibbs JR, Wiedmann M. Molecular subtyping to detect human listeriosis clusters. *Emerging Infectious Diseases*, 9:672-680, 2003. [[online](#)]
54. Odoi A, Martin SW, Michel P, Middleton D, Holt J, Wilson J. Investigation of clusters of giardiasis using GIS and a spatial scan statistic. *International Journal of Health Geographics*, 3:11, 2004. [[online](#)]
55. Jones RC, Liberatore M, Fernandez JR, Gerber SI. Use of a prospective space-time scan statistic to prioritize shigellosis case investigations in an urban jurisdiction. *Public Health Reports*, 121:133-9, 2006.

56. Pearl DL, Louie M, Chui L, Dore K, Grimsrud KM, Leedell D, Martin SW, Michel P, Svenson LW, McEwen SA. The use of outbreak information in the interpretation of clustering of reported cases of *Escherichia coli* O157 in space and time in Alberta, Canada, 2000-2002. *Epidemiology and Infection*, 134:699-711, 2006.
57. de Souza EA, da Silva-Nunes M, Malafronte Rdos S, Muniz PT, Cardoso MA, Ferreira MU. Prevalence and spatial distribution of intestinal parasitic infections in a rural Amazonian settlement, Acre State, Brazil. *Cadernos de Saude Publica*, 23:427-34, 2007. [[online](#)]
58. Osei FB, Duker AA. Spatial dependency of *V. cholera* prevalence on open space refuse dumps in Kumasi, Ghana: a spatial statistical modeling. *International Journal of Health Geographics*, 7:62, 2008. [[online](#)]
59. Sowmyanarayanan TV, Mukhopadhy A, Gladstone BP, Sarkar R, Kang G. Investigation of a hepatitis A outbreak in children in an urban slum in Vellore, Tamil Nadu, using geographic information systems. *Indian Journal of Medical Research*, 128:32-37, 2008.
60. Oviedo M, Munoz P, Dominguez A, Carmona G, Batalla J, Borrás E, Jansá JM. Evaluation of Mass Vaccination Programmes: The experience of hepatitis A in Catalonia (in Spanish). *Revista Española de Salud Pública*, 83:697-709, 2009. [[online](#)]
61. Luquero FJ, Banga CN, Remartínez D, Palma PP, Baron E, Grais RF. Cholera epidemic in Guinea-Bissau (2008): the importance of "place". *PLoS One*, 6:e19005, 2011. [[online](#)]
62. Bompangue Nkoko D, Giraudoux P, Plisnier PD, Tinda AM, Piarroux M, Sudre B, Horion S, Tamfum JJ, Ilunga BK, Piarroux R. Dynamics of cholera outbreaks in Great Lakes region of Africa, 1978-2008. *Emerging Infectious Diseases*, 17:2026-2034, 2011. [[online](#)]

### ***Sexually Transmitted Diseases***

63. Jennings JM, Curriero FC, Celentano D, Ellen JM. Geographic identification of high gonorrhea transmission areas in Baltimore, Maryland. *American Journal of Epidemiology*, 161:73-80, 2005.
64. Wylie JL, Cabral T, Jolly AM. Identification of networks of sexually transmitted infection: a molecular, geographic, and social network analysis. *Journal of Infectious Diseases*, 191:899-906, 2005.
65. Wand H, Ramjee G. Targeting the hotspots: Investigating spatial and demographic variations in HIV infection in small communities in South Africa. *Journal of the International AIDS Society*, 13:41, 2010. [[online](#)]
66. Egger JR, Konty KJ, Borrelli JM, Cumiskey J, Blank S. Monitoring temporal changes in the specificity of an oral HIV test: a novel application for use in postmarketing surveillance. *PLoS One*, 25:e12231, 2010. [[online](#)]
67. Gesink DC, Sullivan AB, Miller WC, Bernstein KT. Sexually transmitted disease core theory: roles of person, place, and time. *American Journal of Epidemiology*. 174:81-9, 2011.

## **Vector Borne Diseases**

68. Fevre EM, Coleman PG, Odiit M, Magona JW, Welburn SC, Woolhouse MEJ. The origins of a new *Trypanosoma brucei rhodesiense* sleeping sickness outbreak in eastern Uganda. *The Lancet*, 358:625-628, 2001.
69. Chaput EK, Meek JJ, Heimer R. Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut. *Emerging Infectious Diseases*, 8:943-948, 2002. [[online](#)]
70. Mostashari F, Kulldorff M, Hartman JJ, Miller JR, Kulasekera V. Dead bird clustering: A potential early warning system for West Nile virus activity. *Emerging Infectious Diseases*, 9:641-646, 2003. [[online](#)]
71. Ghebreyesus TA, Byass P, Witten KH, Getachew A, Haile M, Yohannes M, Lindsay SW. Appropriate Tools and Methods for Tropical Microepidemiology: a Case-study of Malaria Clustering in Ethiopia. *Ethiopian Journal of Health Development*. 17:1-8, 2003.
72. Brooker S, Clarke S, Njagi JK, Polack S, Mugo B, Estambale B, Muchiri E, Magnussen P, Cox J. Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. *Tropical Medicine and International Health*, 9: 757-766, 2004.
73. Washington CH, Radday J, Streit TG, Boyd HA, Beach MJ, Addiss DG, Lovince R, Lovegrove MC, Lafontant JG, Lammie PJ, Hightower AW. Spatial clustering of filarial transmission before and after a Mass Drug Administration in a setting of low infection prevalence. *Filaria Journal*, 3:3, 2004. [[online](#)]
74. Gosselin PL, Lebel G, Rivest S, Fradet MD. The Integrated System for Public Health Monitoring of West Nile Virus (ISPHM-WNV): a real-time GIS for surveillance and decision-making. *International Journal of Health Geographics*, 4:21, 2005. [[online](#)]
75. Gaudart J, Poudiougou B, Ranque S, Doumbo O. Oblique decision trees for spatial pattern detection: optimal algorithm and application to malaria risk. *BMC Medical Research Methodology*, 5:22, 2005. [[online](#)]
76. Nisha V, Gad SS, Selvapandian D, Suganya V, Rajagopal V, Suganti P, Balraj V, Devasundaram J. Geographical information system (GIS) in investigation of an outbreak of dengue fever. *Journal of Communicable Diseases*, 37:39-43, 2005.
77. Reperant LA, Deplazes P. Cluster of *Capillaria hepatica* infections in non-commensal rodents from the canton of Geneva, Switzerland. *Parasitology Research*, 96:340-342, 2005.
78. Fang L, Yan L, Liang S, de Vlas SJ, Feng D, Han X, Zhao W, Xu B, Bian L, Yang H, Gong P, Richardus JH, Cao W. Spatial analysis of hemorrhagic fever with renal syndrome in China. *BMC Infectious Diseases*, 6:77, 2006. [[online](#)]
79. Bonilla RE. Distribución Espacio-Temporal de la Fiebre Dengue en Costa Rica. *Población y Salud en Mesoamérica*, 3:2:2, 2006. [[online](#)]
80. Gaudart J, Poudiougou B, Dicko A, Ranque S, Toure O, Sagara I, Diallo M, Diawara S, Ouattara A, Diakite M, Doumbo OK. Space-time clustering of childhood malaria at the household level: a dynamic cohort in a Mali village. *BMC Public Health*, 6:286, 2006. [[online](#)]



81. Mirghani SE, Nour BY, Bushra SM, Elhassan IM, Snow RW, Noor AM. The spatial-temporal clustering of Plasmodium falciparum infection over eleven years in Gezira State, The Sudan. Malaria Journal, 9:172, 2010. [[online](#)]
82. Haque U, Sunahara T, Hashizume M, Shields T, Yamamoto T, Haque R, Glass GE. Malaria prevalence, risk factors and spatial distribution in a hilly forest area of Bangladesh. PLoS ONE 6(4): e18908, 2011. [[online](#)]
83. Schmidt W-P, Suzuki M, Dinh Thiem V, White RG, Tsuzuki A, Yoshida LM, Yanai H, Haque U, Huu Tho L, Duc Anh D, Ariyoshi K. Population Density, Water Supply, and the Risk of Dengue Fever in Vietnam: Cohort Study and Spatial Analysis. PLoS Medicine, 8:8, e1001082, 2011. [[online](#)]
84. Winskill P, Rowland M, Mtove G, Malima RC, Kirby MJ. Malaria risk factors in north-east Tanzania. Malaria Journal 10:98, 2011. [[online](#)]
85. Washington CH, Radday J, Streit TG, Boyd HA, Beach MJ, Addiss DG, Lovince R, Lovegrove MC, Lafontant JG, Lammie PJ, Hightower AW. Spatial clustering of filarial transmission before and after a mass drug administration in a setting of low infection prevalence. Filaria Journal, 3: 3, 2004. [[online](#)]
86. Bhattarai NR, Van der Auwera G, Rijal S, Picado A, Speybroeck N, Khanal B, De Doncker S, Lal Das M, Ostyn B, Davies C, Coosemans M, Berkvens D, Boelaert M, Dujardin JC. Domestic animals and epidemiology of visceral leishmaniasis, Nepal. Emerging Infectious Diseases, 16:231-237, 2010. [[online](#)]
87. Cook J, Kleinschmidt I, Schwabe C, Nseng G, Bousema T, Corran PH, Riley EM, Drakeley CJ. Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, Equatorial Guinea. PLoS One, 6:e25137, 2011. [[online](#)]
88. Norein AB, Abass MA, Nugud AH, El Hassan I, Snow RW, Noor AM. Identifying residual foci of Plasmodium falciparum infections for malaria elimination: the urban context of Khartoum, Sudan. PLoS One, 6:e16948, 2011. [[online](#)]
89. Rochlin I, Turbow D, Gomez F, Ninivaggi DV, Campbell SR. Predictive mapping of human risk for West Nile virus (WNV) based on environmental and socioeconomic factors. PLoS One. 6:e23280, 2011. [[online](#)]
90. Impoinvil DE, Solomon T, Schluter WW, Rayamajhi A, Bichha RP, Shakya G, Caminade C, Baylis M. The spatial heterogeneity between Japanese encephalitis incidence distribution and environmental variables in Nepal. PLoS One, 6:e22192, 2011. [[online](#)]
91. Bejon P, Turner L, Lavstsen T, Cham G, Olotu A, Drakeley CJ, Lievens M, Vekemans J, Savarese B, Lusingu J, von Seidlein L, Bull PC, Marsh K, Theander TG. Serological evidence of discrete spatial clusters of Plasmodium falciparum parasites. PLoS One, 6:e21711, 2011. [[online](#)]

### ***Other Infectious Diseases***

92. Cousens S, Smith PG, Ward H, Everington D, Knight RSG, Zeidler M, Stewart G, Smith-Bathgate EAB, Macleod MA, Mackenzie J, Will RG. Geographical distribution of variant Creutzfeldt-Jakob disease in Great Britain, 1994-2000. The Lancet, 357:1002-1007, 2001.



93. Huillard d'Aignaux J, Cousens SN, Delasnerie-Laupretre N, Brandel JP, Salomon D, Laplanche JL, Hauw JJ, Alperovitch A. Analysis of the geographical distribution of sporadic Creutzfeldt-Jakob disease in France between 1992 and 1998. *International Journal of Epidemiology*, 31: 490-495, 2002. [[online](#)]
94. Dreesman J, Scharlach H. Spatial-statistical analysis of infectious disease notification data in Lower Saxony. *Gesundheitswesen*, 66: 783-789, 2004.
95. Polack SR, Solomon AW, Alexander NDE, Massae PA, Safari S, Shao JF, Foster A, Mabey DC. The household distribution of trachoma in a Tanzanian village: an application of GIS to the study of trachoma. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 99: 218-225, 2005.

### ***Syndromic Surveillance***

96. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice: The New York City emergency department system. *Emerging Infectious Diseases*, 10:858-864, 2004. [[online](#)]
97. Minnesota Department of Health. Syndromic Surveillance: A New Tool to Detect Disease Outbreaks. *Disease Control Newsletter*, 32:16-17, 2004. [[online](#)]
98. Kleinman K, Abrams A, Kulldorff M, Platt R. A model-adjusted space-time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection*, 2005, 133:409-419.
99. Nordin JD, Goodman MJ, Kulldorff M, Ritzwoller DP, Abrams AM, Kleinman K, Levitt MJ, Donahue J, Platt R. Simulated anthrax attacks and syndromic surveillance. *Emerging Infectious Diseases*, 2005, 11:1394-98. [[online](#)]
100. Yih K, Abrams A, Kleinman K, Kulldorff M, Nordin J, Platt R. Ambulatory-care diagnoses as potential indicators of outbreaks of gastrointestinal illness --- Minnesota. *Morbidity and Mortality Weekly Report*, 54 Suppl:157-62, 2005. [[online](#)]
101. Besculides M, Heffernan R, Mostashari F, Weiss D. Evaluation of school absenteeism data for early outbreak detection, New York City. *BMC Public Health*, 5:105, 2006. [[online](#)]
102. Horst MA, Coco AS. Observing the spread of common illnesses through a community: Using geographic information systems (GIS) for surveillance. *Journal of the American Board of Family Medicine*, 23:32-41, 2010. [[online](#)]
103. van den Wijngaard CC, van Asten L, van Pelt W, Doornbos G, Nagelkerke NJ, Donker GA, van der Hoek W, Koopmans MP. Syndromic surveillance for local outbreaks of lower-respiratory infections: would it work? *PLoS One*, 29:e10406, 2010. [[online](#)]
104. Jones SG, Conner W, Song B, Gordon D, Jayakaran A. Comparing spatio-temporal clusters of arthropod-borne infections using administrative medical claims and state reported surveillance data. *Spatial and Spatio-Temporal Epidemiology*, 2012.

### ***Cancer: Incidence, Prevalence and Mortality***

105. Hjalmar U, Kulldorff M, Gustafsson G, Nagarwalla N. Childhood leukemia in Sweden: Using GIS and a spatial scan statistic for cluster detection. *Statistics in Medicine*, 15:707-715, 1996.

106. Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast cancer in northeastern United States: A geographical analysis. *American Journal of Epidemiology*, 146:161-170, 1997. [[online](#)]
107. Imai J. Spatial disease clustering in Kochi prefecture in Japan. *National Institute of Public Health, Epidemiology and Biostatistics Research*, 57-96, 1998 (in Japanese).
108. VanEenwyk J, Bensley L, McBride D, Hoskins R, Solet D, McKeeman Brown A, Topiwala H, Richter A, Clark R. Addressing community health concerns around SeaTac Airport: Second Report. Washington State Department of Health, 1999. [[online](#)]
109. Hjalmar U, Kulldorff M, Wahlquist Y, Lannering B. Increased incidence rates but no space-time clustering of childhood malignant brain tumors in Sweden. *Cancer*, 85:2077-2090, 1999.
110. Viel JF, Arveux P, Baverel J, Cahn JY. Soft-tissue sarcoma and non-Hodgkin's lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels. *American Journal of Epidemiology*, 152:13-19, 2000.
111. New York State Department of Health. Cancer Surveillance Improvement Initiative, 2001. [[online](#)]
112. Jemal A, Kulldorff M, Devesa SS, Hayes RB, Fraumeni JF. A geographic analysis of prostate cancer mortality in the United States. *International Journal of Cancer*, 101:168-174, 2002.
113. Michelozzi P, Capon A, Kirchmayer U, Forastiere F, Biggeri A, Barca A, Perucci CA. Adult and childhood leukemia near a high-power radio station in Rome, Italy. *American Journal of Epidemiology*, 155:1096-1103, 2002.
114. Zhan FB, Lin H. Geographic patterns of cancer mortality clusters in Texas, 1990 to 1997. *Texas Medicine*, 99:58-64, 2003.
115. Buntinx F, Geys H, Lousbergh D, Broeders G, Cloes E, Dhollander D, Op De Beeck L, Vanden Brande J, Van Waes A, Molenberghs G. Geographical differences in cancer incidence in the Belgian province of Limburg. *European Journal of Cancer*, 39:2058-72, 2003.
116. Santamaria Ulloa C. Evaluación de alarmas por cáncer utilizando análisis espacial: una aplicación para Costa Rica. *Revista Costarricense de Salud Pública*, 12:18-22, 2003. [[online](#)]
117. Sheehan TJ, DeChello LM, Kulldorff M, Gregorio DI, Gershman S, Mrosczyk M. The geographic distribution of breast cancer incidence in Massachusetts 1988-1997, adjusted for covariates. *International Journal of Health Geographics*, 2004, 3:17. [[online](#)]
118. Fang Z, Kulldorff M, Gregorio DI. Brain cancer in the United States, 1986-95: A geographic analysis. *Neuro-Oncology*, 2004, 6:179-187.
119. Hsu CE, Jacobson HE, Soto Mas F. Evaluating the disparity of female breast cancer mortality among racial groups - a spatiotemporal analysis. *International Journal of Health Geographics* 3:4, 2004. [[online](#)]
120. Han DW, Rogerson PA, Nie J, Bonner MR, Vena JE, Vito D, Muti P, Trevisan M, Edge SB, Freudenheim JL. Geographic clustering of residence in early life and subsequent risk of breast cancer (United States). *Cancer Causes and Control*, 15:921-929, 2004.

121. Campo J, Comber H, Gavin A T. All-Ireland Cancer Statistics 1998-2000. Northern Ireland Cancer Registry / National Cancer Registry, 2004. [[online](#)]
122. Hayran M. Analyzing factors associated with cancer occurrence: A geographical systems approach. Turkish Journal of Cancer, 34:67-70, 2004. [[online](#)]
123. Fukuda Y, Umezaki M, Nakamura K, Takano T. Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan. International Journal of Health Geographics, 4:16, 2005. [[online](#)]
124. Ozonoff A, Webster T, Vieira V, Weinberg J, Ozonoff D, Aschengrau A. Cluster detection methods applied to the Upper Cape Cod cancer data. Environmental Health: A Global Access Science Source, 4:19, 2005. [[online](#)]
125. DeChello LM, Sheehan TJ. The geographic distribution of melanoma incidence in Massachusetts, adjusted for covariates. Int J Health Geogr. 2006;5:31 [[online](#)]
126. Gregorio DI, Samociuk H, DeChello L, Swede H. Effects of study area size on geographic characterizations of health events: prostate cancer incidence in Southern New England, USA, 1994–1998. Int J Health Geogr. 5:8, 2006. [[online](#)]
127. Chen Y, Yi Q, Mao Y. Cluster of liver cancer and immigration: a geographic analysis of incidence data for Ontario 1998–2002. Int J Health Geogr. 7:28, 2008. [[online](#)]
128. Lorenzo-Luaces Alvarez P, Guerra-Yi ME, Faes C, Galán Alvarez Y, Molenberghs G. Spatial analysis of breast and cervical cancer incidence in small geographical areas in Cuba, 1999-2003. European Journal of Cancer Prevention, 18:395-403, 2009.
129. Amin R, Bohnert A, Holmes L, Rajasekaran A, Assanasen C. Epidemiologic mapping of Florida childhood cancer clusters. Pediatric Blood Cancer, 54:511-518, 2010.
130. Liu-Mares W, MacKinnon JA, ShermanR, Fleming LE, Rocha-Lima C, Hu, JJ, Lee DJ. Pancreatic cancer clusters and arsenic-contaminated drinking water wells in Florida. BMC Cancer, 13, 111, 2013. [[online](#)]

### ***Cancer: Early versus Late Detection, Stage and Grade***

131. Roche LM, Skinner R, Weinstein RB. 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, 2002.
132. Thomas AJ, Carlin BP. Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering. Statistics in Medicine, 22:113-127, 2003.
133. Sheehan TJ, DeChello LM. A space-time analysis of the proportion of late stage breast cancer in Massachusetts, 1988 to 1997. International Journal of Health Geographics, 4:15, 2005. [[online](#)]
134. Klassen A, Curriero F, Kulldorff M, Alberg AJ, Platz EA, Neloms ST. Missing stage and grade in Maryland prostate cancer surveillance data, 1992-1997. American Journal of Preventive Medicine, 30:S77-87, 2006. [[online](#)]

135. Pollack LA, Gotway CA, Bates JH, Parikh-Patel A, Richards TB, Seeff LC, Hodges H, Kassim S. Use of the spatial scan statistic to identify geographic variations in late stage colorectal cancer in California (United States). *Cancer Causes and Control*, 17:449–457, 2006.
136. DeChello LM, Sheehan TJ. Spatial analysis of colorectal cancer incidence and proportion of late-stage in Massachusetts residents: 1995–1998. *Int J Health Geogr*. 2007;6:20. [[online](#)]

### ***Cancer: Screening, Treatment and Survival***

137. Sheehan TJ, Gershman ST, MacDougal L, Danley R, Mrosszczyk M, Sorensen AM, Kulldorff M. Geographical surveillance of breast cancer screening by tracts, towns and zip codes. *Journal of Public Health Management and Practice*, 6: 48-57, 2001.
138. Gregorio DI, Kulldorff M, Barry L, Samociuk H, Zarfos K. Geographic differences in primary therapy for early stage breast cancer. *Annals of Surgical Oncology*, 2001; 8:844-849, 2001. [[online](#)]
139. Henry KA, Niu X, Boscoe FP. Geographic disparities in colorectal cancer survival. *International Journal of Health Geographics* 2009, 8:48. [[online](#)]

### ***Cardiovascular Diseases***

140. Kuehl KS, Loffredo CA. A cluster of hypoplastic left heart malformation in Baltimore, Maryland *Pediatric Cardiology*, 27:25-31, 2006.
141. Li XY, Chen K. Scan statistic theory and its application in spatial epidemiology (in Chinese). *Zhonghua Liu Xing Bing Xue Za Zhi.*, 29:828-31, 2008.

### ***Rheumatology and Auto-Immune Diseases***

142. Walsh SJ, Fenster JR. Geographical clustering of mortality from systemic sclerosis in the Southeastern United States, 1981-90. *Journal of Rheumatology*, 24:2348-2352, 1997.
143. Walsh SJ, DeChello LM. Geographical variation in mortality from systemic lupus erythematosus in the United States. *Lupus*, 10:637-646, 2001.
144. López-Abente G, Morales-Piga A, Bachiller-Corral FJ, Illera-Martín O, Martín-Domenech R, Abaira V. Identification of possible areas of high prevalence of Paget's disease of bone in Spain. *Clinical and Experimental Rheumatology*, 21:635-368, 2003.
145. Donnan PT, Parratt JDE, Wilson SV, Forbes RB, O'Riordan JI, Swingler RJ. Multiple sclerosis in Tayside, Scotland: detection of clusters using a spatial scan statistic. *Multiple Sclerosis*, 11:403-408, 2005.

### ***Liver Diseases***

146. Ala A, Stanca CM, Bu-Ghanim M, Ahmado I, Branch AD, Schiano TD, Odin JA, Bach N. Increased prevalence of primary biliary cirrhosis near superfund toxic waste sites. *Hepatology*, 43:525-531, 2006.
147. Stanca CM, Babar J, Singal V, Ozdenerol E, Odin JA. Pathogenic role of environmental toxins in immune-mediated liver diseases. *Journal of Immunotoxicology*, 5:59-68, 2008.

148. McNally RJQ, Ducker S, James OFW. Are Transient Environmental Agents Involved in the Cause of Primary Biliary Cirrhosis? Evidence from Space-Time Clustering Analysis. *Hepatology*, 50:1169-1174, 2009.

### ***Diabetes***

149. Green C, Hoppa RD, Young TK, Blanchard JF. Geographic analysis of diabetes prevalence in an urban area. *Social Science and Medicine*, 57:551-560, 2003.
150. Aamodt G, Stene LC, Njølstad PR, Søvik O, Joner G, for the Norwegian Childhood Diabetes Study Group. Spatiotemporal trends and age-period-cohort modelling of the incidence of type 1 diabetes among children ages <15 years in Norway 1973-1982 and 1989-2003. *Diabetes Care*, 30:884-889, 2007.

### ***Allergy and Asthma***

151. Yiannakoulis N, Schopflocher DP, Svenson LW. Using administrative data to understand the geography of case ascertainment. *Chronic Diseases in Canada*, 30:20-28, 2009. [\[online\]](#)

### ***Neurological Diseases***

152. Sabel CE, Boyle PJ, Löytönen M, Gatrell AC, Jokelainen M, Flowerdew R, Maasilta P. Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death. *American Journal of Epidemiology*, 157: 898-905, 2003.

### ***Birth Defects and Other Congenital Outcomes***

153. Kharrazi M, et al. Pregnancy outcomes around the B.K.K. landfill, West Covina, California: An analysis by address. California Department of Health Services, 1998.
154. Bell S. Spatial Analysis of Disease - Applications. In Beam C (ed). *Biostatistical Applications in Cancer Research*. Boston: Kluwer p151-182, 2002. [\[online\]](#)
155. Forand SP, Talbot TO, Druschel C, Cross PK. Data quality and the spatial analysis of disease rates: congenital malformations in New York State. *Health and Place*, 8:191-199, 2002.
156. Colorado Department of Public Health and Environment. Analysis of birth defect data in the vicinity of the Redfield plume area in southeastern Denver county: 1989-1999. Colorado Department of Public Health and the Environment, 2002. [\[online\]](#)
157. Boyle E, Johnson H, Kelly A, McDonnell R. Congenital anomalies and proximity to landfill sites. *Irish Medical Journal*, 97:16-18, 2004.
158. Ozdenerol E, Williams BL, Kang SY, Magsumbol MS. Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters. *International Journal of Health Geographics*, 4:19, 2005. [\[online\]](#)
159. Viel JF, Floret N, Mauny F. Spatial and space-time scan statistics to detect low clusters of sex ratio. *Environmental and Ecological Statistics*, 12:289-299, 2005.
160. Grady SC, Enander H. Geographic analysis of low birthweight and infant mortality in Michigan using automated zoning methodology *International Journal of Health Geographics* 2009, 8:10. [\[online\]](#)

## ***Pediatrics***

161. George M, Wiklund L, Aastrup M, Pousette J, Thunholm B, Saldeen T, Wernroth L, Zaren B, Holmberg L. Incidence and geographical distribution of sudden infant death syndrome in relation to content of nitrate in drinking water and groundwater levels. *European Journal of Clinical Investigation*, 31: 1083-1094, 2001.
162. Sankoh OA, Ye Y, Sauerborn R, Muller O, Becher H. Clustering of childhood mortality in rural Burkina Faso. *International Journal of Epidemiology*, 30:485-492, 2001. [[online](#)]
163. Ali M, Asefaw T, Byass P, Beyene H, Karup Pedersen F. Helping northern Ethiopian communities reduce childhood mortality: population-based intervention trial. *Bulletin of the World Health Organization*. 83:27-33, 2005. [[online](#)]
164. Awini E, Mattah P, Sankoh O, Gyapong M. Spatial variations in childhood mortalities at the Dodowa Health and Demographic Surveillance System site of the INDEPTH Network in Ghana. *Tropical Medicine and International Health*, 2010.

## ***Geriatrics***

165. Yiannakoulis N, Rowe BH, Svenson LW, Schopflocher DP, Kelly K, Voaklander DC. Zones of prevention: the geography of fall injuries in the elderly. *Social Science and Medicine*, 57:2065-73, 2003.
166. Vaneckova P, Beggs PJ, Jacobson CR. Spatial analysis of heat-related mortality among the elderly between 1993 and 2004 in Sydney, Australia. *Social Science and Medicine*, 70:293-304, 2010.

## ***Psychology***

167. Margai F, Henry N. A community-based assessment of learning disabilities using environmental and contextual risk factors. *Social Science and Medicine*, 56: 1073-1085, 2003.

## ***Brain Imaging***

168. Yoshida M, Naya Y, Miyashita Y. Anatomical organization of forward fiber projections from area TE to perirhinal neurons representing visual long-term memory in monkeys. *Proceedings of the National Academy of Sciences of the United States of America*, 100:4257-4262, 2003. [[online](#)]

## ***Alcohol and Drugs***

169. Hanson CE, Wiczorek WF. Alcohol mortality: a comparison of spatial clustering methods. *Social Science and Medicine*, 55:791-802, 2002.

## ***Accidents and Suicide***

170. Nkhoma ET, Hsu CE, Hunt VI, Harris AM. Detecting spatiotemporal clusters of accidental poisoning mortality among Texas counties, U.S., 1980 - 2001. *International Journal of Health Geographics*, 3:25, 2004. [[online](#)]
171. Exeter DJ, Boyle PJ. Does young adult suicide cluster geographically in Scotland? *Journal of Epidemiology and Community Health*, 61:731-736, 2007.



172. Warden R. Comparison of Poisson and Bernoulli spatial cluster analyses of pediatric injuries in a fire district. *International Journal of Health Geographics*, 7:51, 2008. [[online](#)]
173. Mesoudi A. The cultural dynamics of copycat suicide. *PLoS One*, 4:e7252, 2009. [[online](#)]
174. Saman DM, Cole HP, Odoi A, Myers ML, Carey DI, Westneat SC. A spatial cluster analysis of tractor overturns in Kentucky from 1960 to 2002. *PLoS One*, 7:e30532, 2012. [[online](#)]
175. Amin R, Ritter EK, Cossette L. A Geospatial Analysis of Shark Attack Rates for the Coast of California: 1994–2010. *Journal of Environment and Ecology*, 3:246-255, 2012.
176. Fuchs S, Ornetsmüller C, Totschnig R. Spatial scan statistics in vulnerability assessment – an application to mountain hazards. *Natural Hazards* 64:2129-2151, 2012.
177. Campo J. Firearm deaths in Washington State. Washington State Health Services Research Brief No. 71, 2013. [[online](#)]

### ***Demography***

178. Collado Chaves A. Fecundidad adolescente en el gran área metropolitana de Costa Rica. *Población y Salud en Mesoamérica*, 1:4, 2003. [[online](#)]

### ***Veterinary Medicine, Domestic Animals***

179. Norström M, Pfeiffer DU, Jarp J. A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds. *Preventive Veterinary Medicine*, 47: 107-119, 2000.
180. Ward MP. Blowfly strike in sheep flocks as an example of the use of a time-space scan statistic to control confounding. *Preventive Veterinary Medicine*, 49: 61-69, 2001.
181. United States Department of Agriculture. West Nile virus in equids in the Northeastern United States in 2000. USDA, APHIS, Veterinary Services, 2001. [[online](#)]
182. Doherr MG, Hett AR, Rufenacht J, Zurbriggen A, Heim D. Geographical clustering of cases of bovine spongiform encephalopathy (BSE) born in Switzerland after the feed ban. *Veterinary Record*, 151: 467-472, 2002.
183. Perez AM, Ward MP, Torres P, Ritacco V. Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina. *Preventive Veterinary Medicine*, 56: 63-74, 2002.
184. Schwermer H, Rufenacht J, Doherr MG, Heim D. Geographic distribution of BSE in Switzerland. *Schweizer Archiv für Tierheilkunde*, 144:701-708, 2002.
185. Ward MP. Clustering of reported cases of leptospirosis among dogs in the United States and Canada. *Preventive Veterinary Medicine*, 56:215-226, 2002.
186. Falconi F, Ochs H, Deplazes P. Serological cross-sectional survey of psoroptic sheep scab in Switzerland. *Veterinary Parasitology*, 109:119-127, 2002.
187. Knuesel R, Segner H, Wahli T. A survey of viral diseases in farmed and feral salmonids in Switzerland. *Journal of Fish Diseases*, 26:167-182, 2003.

188. Berke O, Grosse Beilage E. Spatial relative risk mapping of pseudorabies-seropositive pig herds in an animal-dense region. *Journal of Veterinary Medicine*, B50: 322–325, 2003.
189. Abrial D, Calavas D, Lauvergne N, Morignat E, Ducrot C. Descriptive spatial analysis of BSE in western France. *Veterinary Research*, 34:749-60, 2003.
190. Moore GE, Ward MP, Kulldorff M, Caldanaro RJ, Guptill LF, Lewis HB, Glickman LT. A space-time cluster of adverse events associated with a canine rabies vaccine. *Vaccine*, 23:5557-62, 2005.
191. Sheridan HA, McGrath G, White P, Fallon R, Shoukri MM, Martin SW. A temporal-spatial analysis of bovine spongiform encephalopathy in Irish cattle herds, from 1996 to 2000. *Canadian Journal of Veterinary Research*, 69:19-25, 2005. [[online](#)]
192. Guerin MT, Martin SW, Darlington GA, Rajic A. A temporal study of *Salmonella* serovars in animals in Alberta between 1990 and 2001. *Canadian Journal of Veterinary Research*, 69:88-89, 2005. [[online](#)]
193. Allepuz A, López-Quílez A, Forte A, Fernández G, Casal J. Spatial analysis of bovine spongiform encephalopathy in Galicia, Spain (2000-2005). *Preventive Veterinary Medicine*, 79:174-85, 2007.
194. Heres L, Brus DJ, Hagenaars TJ. Spatial analysis of BSE cases in the Netherlands. *BMC Veterinary Research*, 4:21, 2008. [[online](#)]
195. Frossling J, Nodtvedt A, Lindberg A, Björkman C. Spatial analysis of *Neospora caninum* distribution in dairy cattle from Sweden. *Geospatial Health*, 3:39-45, 2008.

### ***Veterinary Medicine, Wildlife***

196. Smith KL, DeVos V, Bryden H, Price LB, Hugh-Jones ME, Keim P. *Bacillus anthracis* diversity in Kruger National Park. *Journal of Clinical Microbiology*, 38:3780-3784, 2000. [[online](#)]
197. Berke O, von Keyserlingk M, Broll S, Kreienbrock L. On the distribution of *Echinococcus multilocularis* in red foxes in Lower Saxony: identification of a high risk area by spatial epidemiological cluster analysis. *Berliner und Münchener Tierärztliche Wochenschrift*. 115:428-434, 2002.
198. Miller MA, Gardner IA, Kreuder C, Paradies DM, Worcester KR, Jessup DA, Dodd E, Harris MD, Ames JA, Packham AE, Conrad PA. Coastal freshwater runoff is a risk factor for *Toxoplasma gondii* infection of southern sea otters (*Enhydra lutris nereis*). *International Journal for Parasitology*, 32:997-1006, 2002.
199. Hoar BR, Chomel BB, Rolfe DL, Chang CC, Fritz CL, Sacks BN, Carpenter TE. Spatial analysis of *Yersinia pestis* and *Bartonella vinsonii* subsp *berkhoffii* seroprevalence in California coyotes (*Canis latrans*). *Preventive Veterinary Medicine*, 56:299-311, 2003.
200. Olea-Popelka FJ, Griffin JM, Collins JD, McGrath G, Martin SW. Bovine tuberculosis in badgers in four areas in Ireland: does tuberculosis cluster? *Preventive Veterinary Medicine*, 59:103-111, 2003.
201. Joly DO, Ribic CA, Langenberg JA, Beheler K, Batha CA, Dhuey BJ, Rolley RE, Bartelt G, Van Deelen TR, Samuel MD. Chronic wasting disease in free-ranging Wisconsin white-tailed deer. *Emerging Infectious Disease*, 9: 599-601, 2003. [[online](#)]



202. Miller MA, Grigg ME, Kreuder C, James ER, Melli AC, Crosbie PR, Jessup DA, Boothroyd JC, Brownstein D, Conrad PA. An unusual genotype of *Toxoplasma gondii* is common in California sea otters (*Enhydra lutris nereis*) and is a cause of mortality. *International Journal for Parasitology*, 34:275-284, 2004.
203. Olea-Popelka FJ, Flynn O, Costello E, McGrath G, Collins JD, O’Keeffe JO, Kelton DF, Berke O, Martin SW. Spatial relationship between *Mycobacterium bovis* strains in cattle and badgers in four areas in Ireland. *Preventive Veterinary Medicine*, 71:57-70, 2005.

### **Mammalogy**

204. Webb NF, Hebblewhite M, Merrill EH. Statistical Methods for Identifying Wolf Kill Sites Using Global Positioning System Locations. *Journal of Wildlife Management*, 2008, 72, 798-807.
205. McPhee HM, Webb NF, Merrill EH. Hierarchical predation: Wolf (*Canis lupus*) selection along hunt paths and at kill sites. *Canadian Journal of Zoology*, 2012, 90:555-563.

### **Entomology**

206. Porcasi X, Catalá SS, Hrellac H, Scavuzzo MC, Gorla DE. Infestation of Rural Houses by *Triatoma Infestans* (Hemiptera: Reduviidae) in Southern Area of Gran Chaco in Argentina. *Journal of Medical Entomology*, 43:1060-1067, 2006.

### **Ichthyology**

207. Spindler BD, Chipps SR, Klumb RA, Wimberly MC. Spatial analysis of pallid sturgeon *Scaphirhynchus albus* distribution in the Missouri River, South Dakota. *Journal of Applied Ichthyology*, 25:8-13, 2009.

### **Botany**

208. Bayon C, Pei MH, Ruiz C, Hunter T. Genetic structure and spatial distribution of the mycoparasite *Sphaerellopsis filum* on *Melampsora larici-epitea* in a short-rotation coppice willow planting. *Plant Pathology*, 56:616-623, 2007.

### **Forestry**

209. Coulston JW, Riitters KH. Geographic analysis of forest health indicators Using Spatial Scan Statistics. *Environmental Management*, 31: 764-773, 2003.
210. Riitters KH, Coulston JW. Hot spots of perforated forest in the eastern United States. *Environmental Management*, 35:483-492, 2005.
211. Tuia D, Ratle F, Lasaponara R, Telesca L, Kanevski M. Scan statistics analysis of forest fire clusters. *Communications in Nonlinear Sciences and Numerical Simulations*, 13:1689-94, 2008.
212. Tonini M, Tuia D, Ratle F. Detection of clusters using space–time scan statistics. *International Journal of Wildland Fire*, 18 830–836, 2009.
213. Fei S. Applying hotspot detection methods in forestry: A case study of Chestnut Oak regeneration. *International Journal of Forestry Research*., 815292, 2010. [[online](#)]

214. Vega Orozco C, Tonini M, Conedera M, Kanveski M. Cluster recognition in spatial-temporal sequences: the case of forest fires. *Geoinformatica*, 16: 653-673, 2012. [[online](#)]

### ***Environment***

215. Vadrevu KP. Analysis of fire events and controlling factors in eastern India using spatial scan and multivariate statistics. *Geografiska Annaler*, 90A: 315-328, 2008.
216. Sudakin DL, Horowitz Z, Giffin S. Regional variation in the incidence of symptomatic pesticide exposures: Applications of geographic information systems. *Journal of Toxicology - Clinical Toxicology*, 40:767-773, 2002.

### ***Natural Disasters***

217. Witham CS, Oppenheimer C. Mortality in England during the 1783-4 Laki Craters eruption. *Bulletin of Volcanology*, 67:15-25, 2004.
218. Stevenson JR Emrich CT Mitchell JT, Cutter SL. Using building permits to monitor disaster recovery: A spatio-temporal case study of coastal Mississippi following hurricane Katrina, *Cartography and Geographic Information Science*, 37:S57-68, 2010.

### ***War***

219. Ziemke J. From battles to massacres. 3rd Annual Harvard-Yale-MIT Graduate Student Conference on Order, Conflict and Violence, 2008. [[online](#)]
220. O'Loughlin J, Witmer F, Linke A. The Afghanistan-Pakistan Wars 2008–2009: Micro-geographies, Conflict Diffusion, and Clusters of Violence. *Eurasian Geography and Economics*, 2010, 51, 437-71. [[online](#)]
221. O'Loughlin J, Witmer FDW, Linke AM, Thorwardson N. Peering into the Fog of War: The Geography of the WikiLeaks Afghanistan War Logs, 2004–2009. *Eurasian Geography and Economics*, 51:472–495, 2010. [[online](#)]
222. O'Loughlin J, Witmer FDW, The Localized Geographies of Violence in the North Caucasus of Russia, 1999-2007', *Annals of the Association of American Geographers*, 101: 178-201, 2011. [[online](#)]

### ***Criminology***

223. Jefferis ES. A multi-method exploration of crime hot spots: SaTScan results. National Institute of Justice, Crime Mapping Research Center, 1998.
224. Kaminski RJ, Jefferis ES, Chanhatisilpa C. A spatial analysis of American police killed in the line of duty. In Turnbull et al. (eds.), *Atlas of crime: Mapping the criminal landscape*. Phoenix, AZ: Oryx Press, 2000.
225. LeBeau JL. Demonstrating the analytical utility of GIS for police operations: A final report. National Criminal Justice Reference Service, 2000. [[online](#)]

226. Beato Filho CC, Assunção RM, Silva BF, Marinho FC, Reis IA, Almeida MC. Homicide clusters and drug traffic in Belo Horizonte, Minas Gerais, Brazil from 1995 to 1999. *Cadernos de Saúde Pública*, 17:1163-1171, 2001. [[online](#)]
227. Ceccato V, Haining R. Crime in border regions: The Scandinavian case of Öresund, 1998-2001. *Annals of the Association of American Geographers*, 94:807-826, 2004.
228. Ceccato V. Homicide in Sao Paulo, Brazil: Assessing the spatial-temporal and weather variations. *Journal of Environmental Psychology*, 25:307-321, 2005.
229. Minamisava R, Nouer SS, de Moraes Neto OL, Melo LK, Andrade ALS. Spatial clusters of violent deaths in a newly urbanized region of Brazil: Highlighting the social disparities. *International Journal of Health Geographics*, 8:66, 2009. [[online](#)]
230. Nakaya T, Yano K. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14:223-239, 2010.
231. Leitner M, Helbich M. The Impact of Hurricanes on Crime: A Spatio-temporal Analysis in the City of Houston, TX. *Cartography and Geographic Information Science*, 37:214-222, 2011.

### ***Urban and Rural Planning***

232. Huang L, Stinchcomb DG, Pickle LW, Dill J, Berrigan D. Identifying clusters of active transportation using spatial scan statistics. *American Journal of Preventive Medicine*, 37:157-166, 2009.
233. Helbich M. Beyond potsuburbia? Multifunctional service agglomeration in Vienna's urban fringe. *Journal of Economic and Social Geography*, 2011.

### ***History and Archeology***

234. Usher BM, Allen KL. Identifying kinship clusters: SaTScan for genetic spatial analysis. *American Journal of Physical Anthropology, Supplement*, 126,S40,210, 2005.
235. Wang F, Hartmann J, Luo W, Huang P. GIS-based spatial analysis of Tai place names in southern China: An exploratory study of methodology. *Annals of GIS*, 12:1-9, 2006. [[online](#)]

### ***Astronomy***

236. Marcos RDLF, Marcos CDLF. From star complexes to the field: Open cluster families, 672:342-351, 2008.
237. Bidin CM, Marcos RD, Marcos CD, Carraro, G. Not an open cluster after all: the NGC 6863 asterism in Aquila. *Astronomy and Astrophysics*, 510:A44, 2010.[[online](#)]

**Related Topics:** *Methodological Papers, SaTScan Bibliography, Suggested Citation.*

---

## Other References Mentioned in the User Guide

- 238. Alt KW, Vach W. The reconstruction of 'genetic kinship' in prehistoric burial complexes - problems and statistics. In Bock HH, Ihm P (eds): Classification, data analysis, and knowledge organization. Berlin: Springer Verlag, 1991.
- 239. Baker RD. Testing for space-time clusters of unknown size. *Journal of Applied Statistics*, 23:543-554, 1996.
- 240. Besag J, Newell J. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society*, A154:143-155, 1991.
- 241. Bithell JF. The choice of test for detecting raised disease risk near a point source. *Statistics in Medicine*, 14:2309-2322, 1995.
- 242. Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society*, B52:73-104, 1990.
- 243. Diggle PJ, Chetwynd AD. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, 47:1155-1163, 1991.
- 244. Diggle P, Chetwynd AG, Häggkvist R, Morris SE. Second-order analysis of space-time clustering. *Statistical Methods in Medical Research*, 4:124-136, 1995.
- 245. Glaz J, Balakrishnan N (editors). *Scan Statistics and Applications*. Birkhäuser: Boston, 1999.
- 246. Glaz J, Naus JI, Wallenstein S. *Scan Statistics*. Springer Verlag: New York, 2001.
- 247. Glaz J, Pozdnyakov V, Wallenstein S. *Scan Statistics: Theory and Applications*. Birkhäuser: Boston, 2009.
- 248. Grimson RC. A versatile test for clustering and a proximity analysis of neurons. *Methods of Information in Medicine*, 30:299-303, 1991.
- 249. Jacquez GM. A k nearest neighbor test for space-time interaction. *Statistics in Medicine*, 15:1935-1949, 1996.
- 250. Knox G. The detection of space-time interactions. *Applied Statistics*, 13:25-29, 1964.
- 251. Kulldorff M. Statistical Methods for Spatial Epidemiology: Tests for Randomness, in *GIS and Health in Europe*, Löytönen M and Gatrell A (eds), London: Taylor & Francis, 1998.
- 252. Kulldorff M, Hjalmar U. The Knox method and other tests for space time interaction. *Biometrics*, 9:621-630, 1999.
- 253. Lawson AB. On the analysis of mortality events associated with a pre-specified fixed point. *Journal of the Royal Statistical Society, Series A*, 156:363-377, 1993.
- 254. Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27:209-220, 1967.

255. Moran PAP. Notes on continuous stochastic phenomena. *Biometrika*, 37:17-23, 1950.
256. Naus J. The distribution of the size of maximum cluster of points on the line. *Journal of the American Statistical Association*, 60:532-538, 1965.
257. Openshaw S, Charlton M, Wymer C, Craft AW: A mark 1 analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, 1, 335-358, 1987.
258. Ranta J, Pitkniemi J, Karvonen M, et al. Detection of overall space-time clustering in non-uniformly distributed population. *Statistics in Medicine*, 15:2561-2572, 1996.
259. Rushton G, Lolonis P. Exploratory Spatial Analysis of Birth Defect Rates in an Urban Population. *Statistics in Medicine*, 7:717-726, 1996.
260. Stone RA. Investigation of excess environmental risk around putative sources: statistical problems and a proposed test. *Statistics in Medicine*, 7:649-660, 1988.
261. Tango T. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine*, 14:2323-2334, 1995.
262. Tango T. A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine*, 19:191-204, 2000.
263. Turnbull B, Iwano EJ, Burnett WS, et al. Monitoring for clusters of disease: application to Leukemia incidence in upstate New York. *American Journal of Epidemiology*, 132:S136-143, 1990.
264. Waller LA, Turnbull BW, Clark LC, Nasca P. Chronic disease surveillance and testing of clustering of disease and exposure. *Environmetrics*, 3:281-300, 1992.
265. Walter SD. A simple test for spatial pattern in regional health data. *Statistics in Medicine*, 13:1037-1044, 1994.
266. Whittemore AS, Friend N, Brown BW, Holly EA. A test to detect clusters of disease. *Biometrika*, 74:631-635, 1987.