Multinomial Scan Statistic for Identifying Unusual Population Age Structures

Francis Boscoe, New York State Department of Health

Martin Kulldorff, Brigham and Women's Hospital, Harvard Medical School

March, 2018

1. Introduction

SaTScan^{TM} is free software that analyzes spatial, temporal, and space-time data using scan statistics. It is designed to detect spatial or space-time disease clusters, and to determine if they are statistically significant. The software may also be used for similar problems in other fields such as archaeology, criminology, demography, ecology, geography or zoology. A wide list of published application areas can be found in the SaTScan bibliography: http://www.satscan.org/references.html

This is a step-by-step tutorial for using SaTScan software to analyze unordered categorical data. The tutorial can be utilized independently, without first having worked through other SaTScan tutorials. It was written using SaTScan version 9.5 for Windows, released in January, 2018. The software tabs for later versions may be slightly different than the screen shots shown in this tutorial, but they will be nearly the same and there should not be a problem using the tutorial with later versions. This tutorial can also be used with the Linux and Apple IOS versions of SaTScan, except that some of the file handling steps will have to be adapted to those operating systems.

This tutorial is intended for self-learning, but it can also be used in a classroom setting. We recommend using it as a complement to the SaTScan User Guide and to the various scientific publications describing the statistical methods. The only prerequisite knowledge is a basic understanding of statistics and epidemiology.

In this tutorial, we use the purely spatial scan statistic to analyze the geographical variation in age-specific populations in the United States in order to determine if there are any geographical clusters of populations which are unusually young, old, middle-aged, or any combination of these.

2. What is the Multinomial Model?

SaTScan contains a number of different methods for detecting clusters. The most commonly used methods divide the data into binary categories: people with and without disease, victims and non-victims of crime, grave sites with and without persons with altered wisdom teeth, and so on. In this tutorial, we use a method which categorizes data into three or more categories, known as the multinomial model. This method considers all possible groupings of the categories and looks for instances where a grouping within a geographic area is different than that outside the area. For example, we might be interested in studying cases of meningitis which are grouped into five different clonal complexes. The multinomial method will look for geographic clusters of meningitis overall, within each of the five clonal complexes, and within all combinations of complexes.

For this tutorial, we chose to use populations by five-year age group for counties in the United States, as measured in the 2010 United States census. We chose this data set because it is freely available and because most users have prior knowledge that will make the data set more interesting to work with. For example, most of us are aware that the population skews older in South Florida and younger in the Pacific Northwest. Will SaTScan confirm these expectations? What other unusual age distributions might we be less familiar with? This is not purely an exercise in recreational geography, as these findings can help inform decisions about allocation of social services, business location decisions, marketing campaigns, and so on.

3. United States Population Data

2010 populations categorized by age group for each county in the United States were obtained from the Census Bureau and saved in the file populations_by_age.cas. The file consists of 4 fields: state and county name, 5-digit state and county code, population, and age group. Age groups are numbered 1-18 where 1 is the youngest group (ages 0-4), 2 the next youngest (ages 5-9), and so on, through 17 (ages 80-84) and 18 (ages 85+). The file includes 50 states and the District of Columbia but excludes Puerto Rico and island territories.

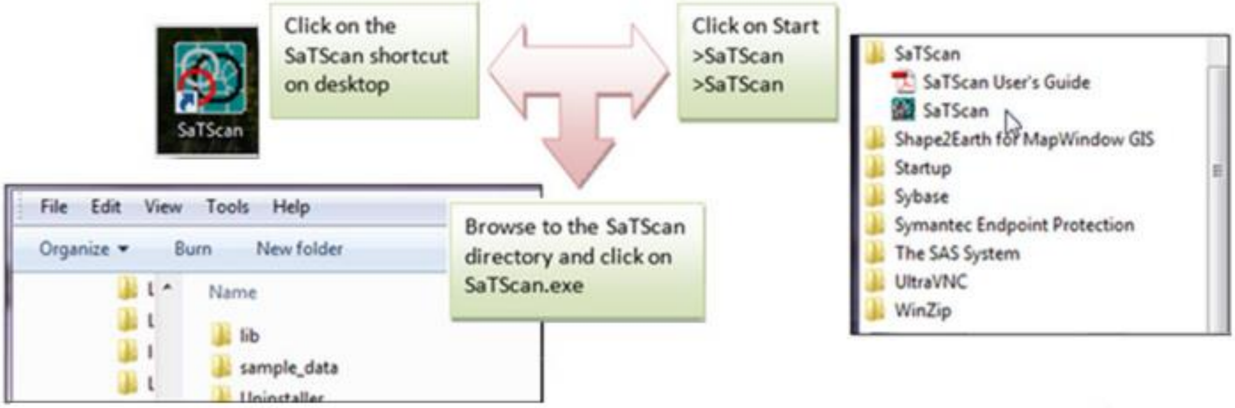
4. SaTScan Software Download and Installation

To download the free SaTScan software, please go to https://www.satscan.org. Select “download”, and follow the instructions. To obtain the download password, you need to register, providing your name, email address, organizational affiliation and country.

The SaTScan software can depict clusters with Google Earth. This requires Google Earth to be installed on your computer. To download, go to https://www.google.com/earth/index.html, and follow the instructions. This is an optional step; SaTScan can be run without Google Earth.

5. Launching the SaTScan Software

Launch the SaTScan software by using one of the following three methods. Your setup may vary slightly on your computer.

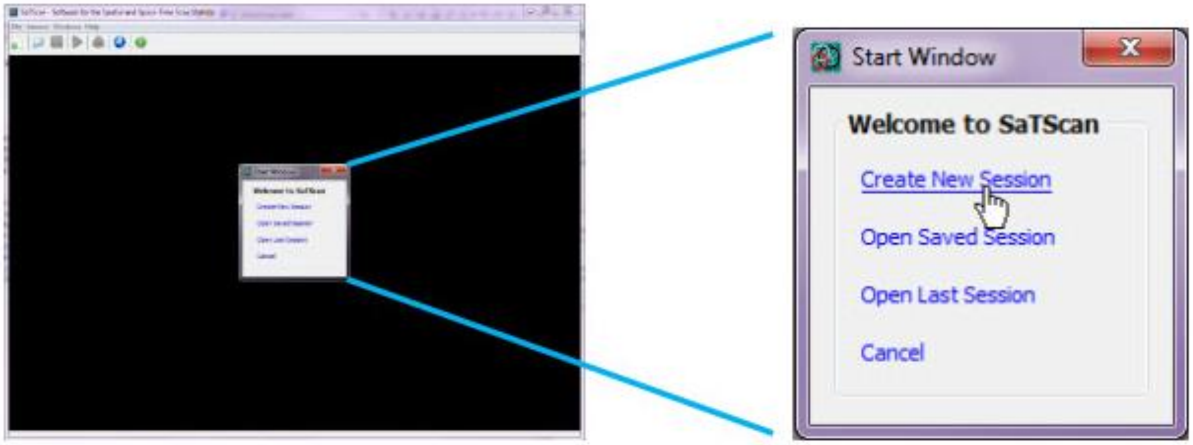


Click on the SaTScan shortcut on desktop

Click on Start >SaTScan >SaTScan

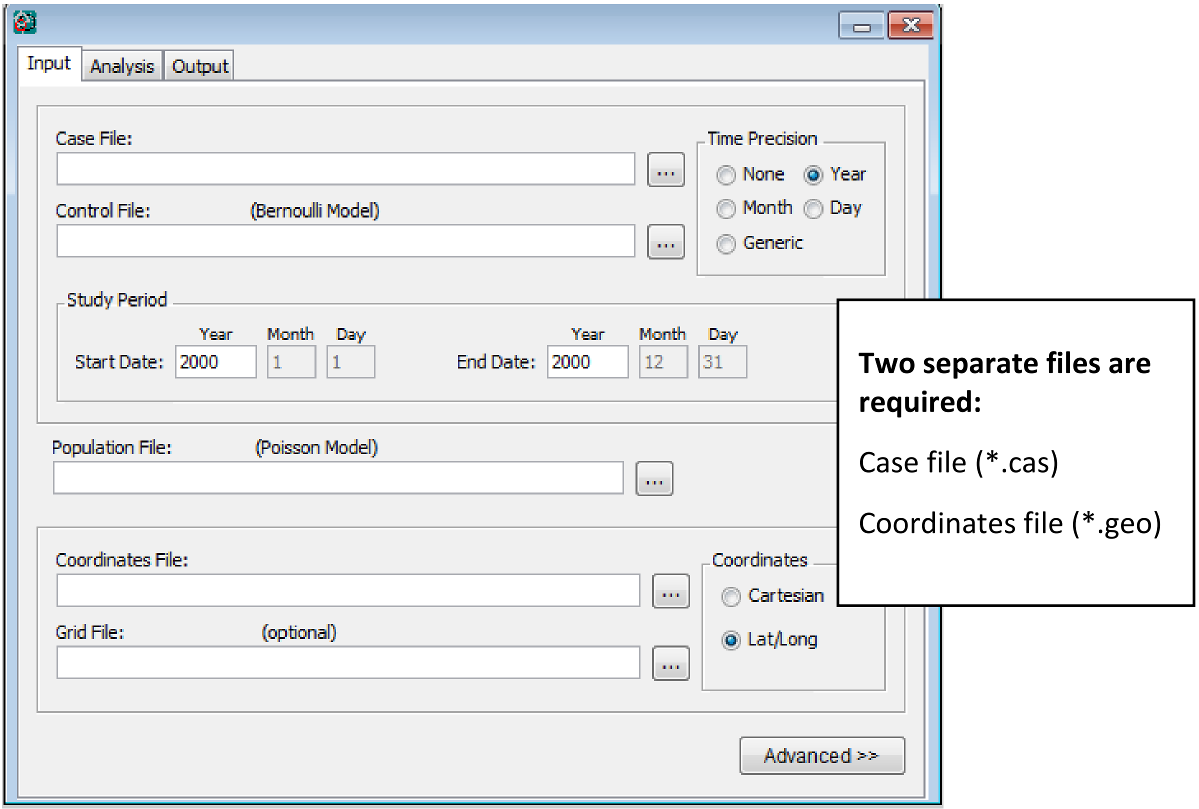
Browse to the SaTScan directory and click on SaTScan.exe

Next, select “Create New Session” from the Start Window.



You should now see the input data tab, and you are ready to specify your analysis parameters. The SaTScan software has three main tabs: input, analysis, and output. We will go over each in turn.

6. Input Tab



Case File:

Control File: (Bernoulli Model)

Time Precision: ○ None ● Year ○ Month ○ Day ○ Generic

Study Period

	Year	Month	Day
Start Date:	2000	1	1
End Date:	2000	12	31

Population File: (Poisson Model)

Coordinates File:

Grid File: (optional)

Coordinates: ○ Cartesian ● Lat/Long

Two separate files are required:

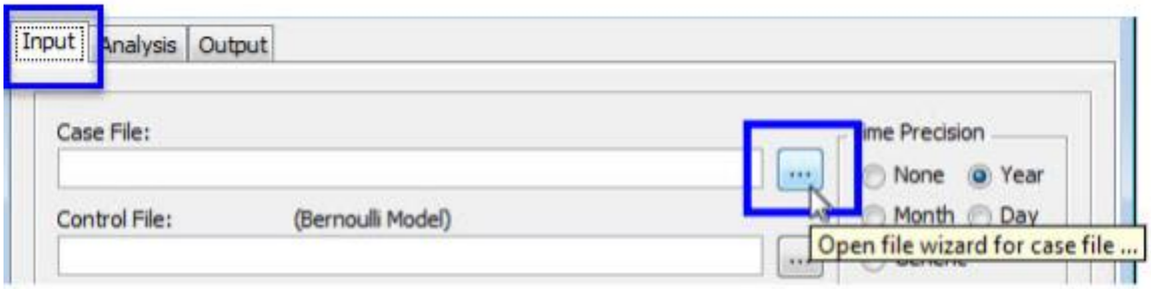
Case file (*.cas)

Coordinates file (*.geo)

The first of the three main tabs is used to specify the input data. When using the Multinomial model, two input files are required. The first is a “case” file, so named because SaTScan is most often used with disease cases; here the “cases” are the populations by age group. The second is a coordinates file containing the latitude and longitude coordinates. The two files are named populations_by_age.cas and coordinates.geo, respectively

6.1 Case File (*.cas)

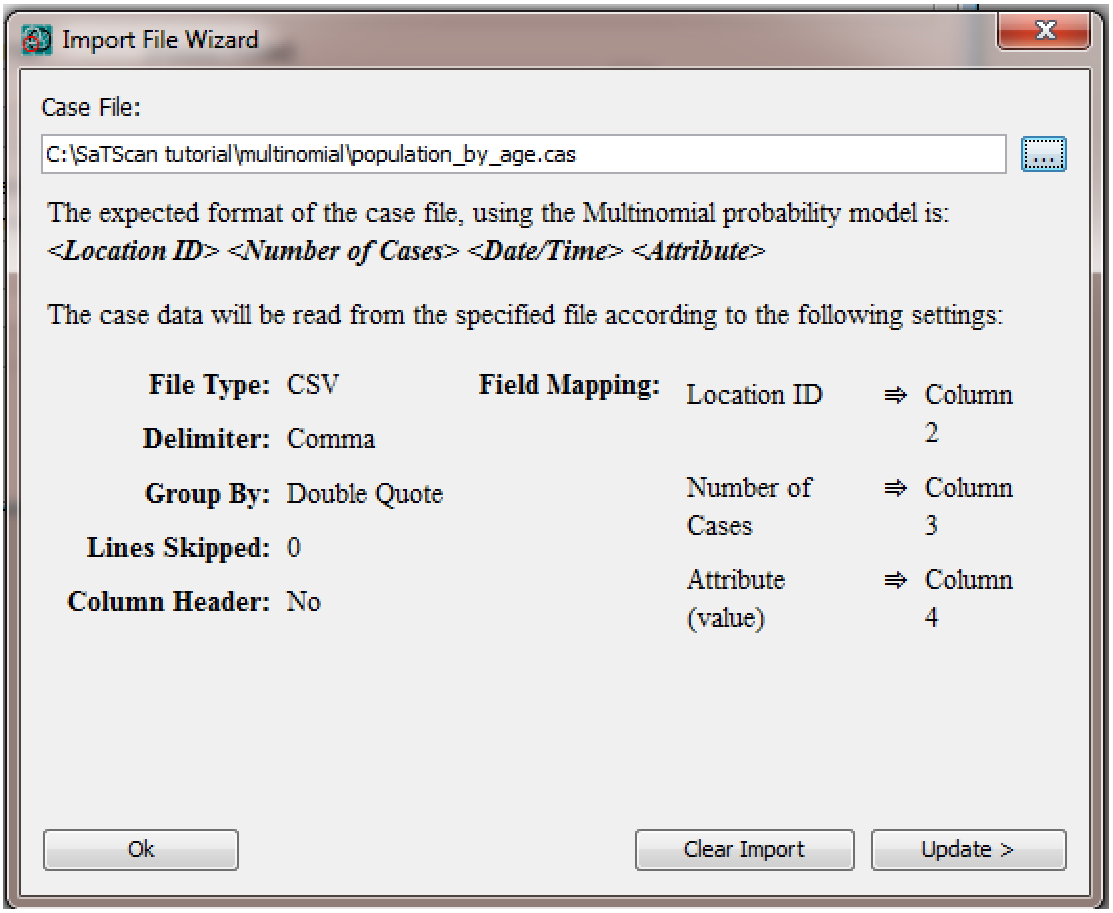
We will use the SaTScan Import Wizard to add the case file.



Click on the [...] **button** to import the case data using the SaTScan file wizard.

Select population_by_age.cas as the import file.

The .cas file is a comma-separated text file. Opening this file will bring you to the Import File Wizard screen:



Case File:

C:\SaTScan tutorial\multinomial\population_by_age.cas

The expected format of the case file, using the Multinomial probability model is:
<Location ID> <Number of Cases> <Date/Time> <Attribute>

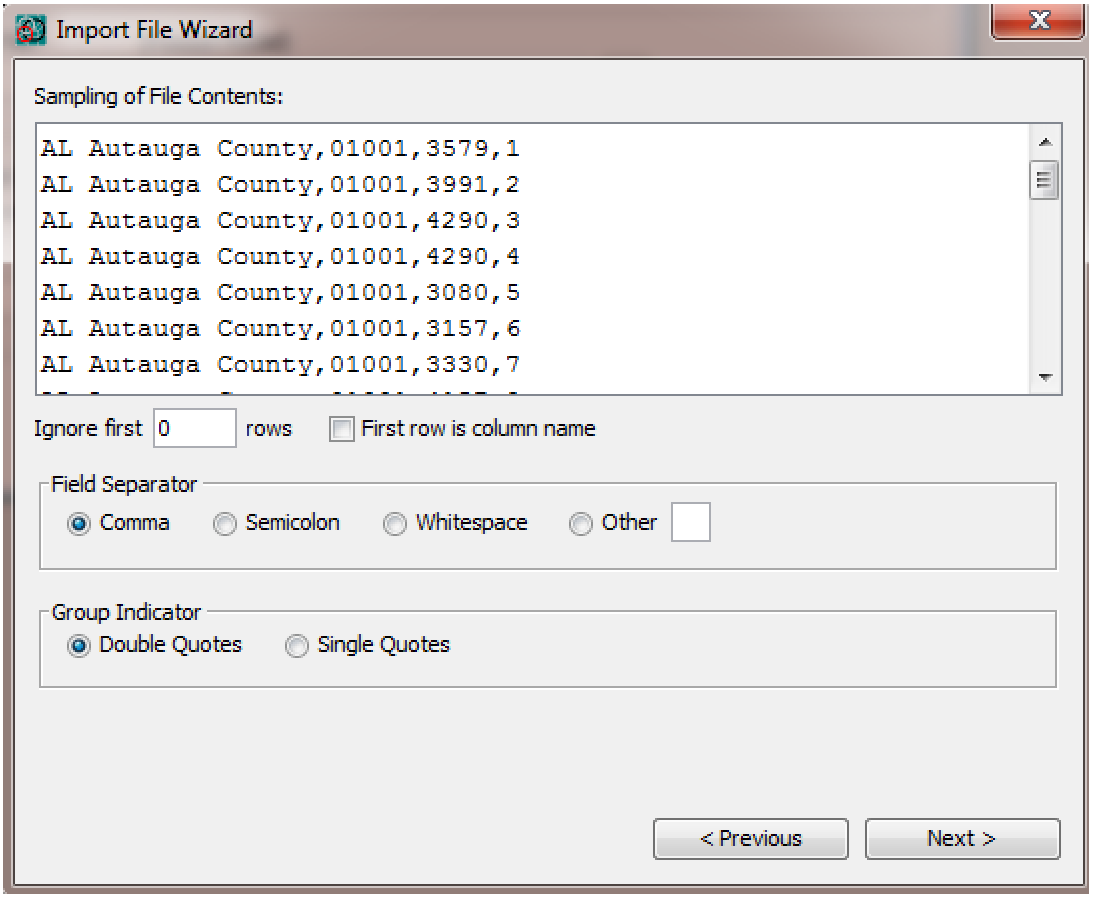
The case data will be read from the specified file according to the following settings:

File Type: CSV
Delimiter: Comma
Group By: Double Quote
Lines Skipped: 0
Column Header: No

Field Mapping:

Location ID	⇛	Column 2
Number of Cases	⇛	Column 3
Attribute (value)	⇛	Column 4

Because the file is not in the expected format, we need to tell the software which columns are which, by hitting the Update button.



Sampling of File Contents:

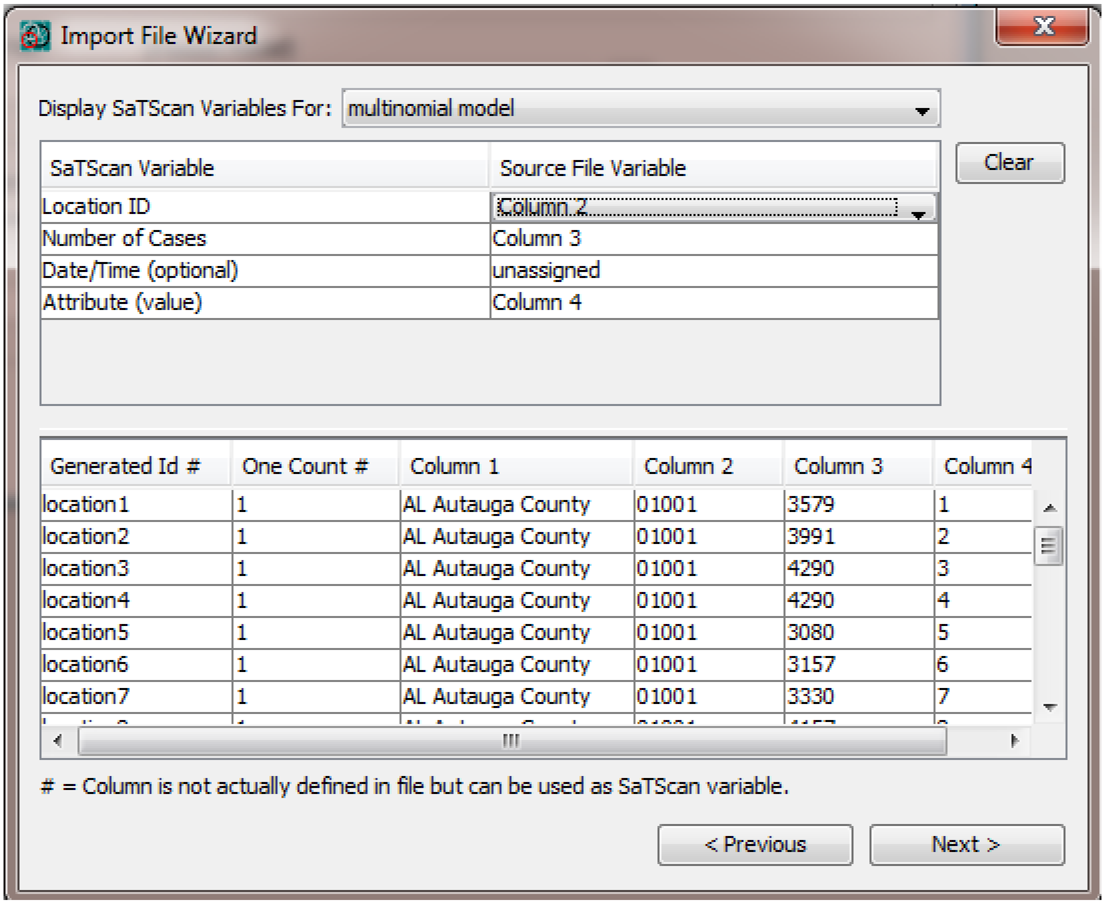
```
AL Autauga County,01001,3579,1
AL Autauga County,01001,3991,2
AL Autauga County,01001,4290,3
AL Autauga County,01001,4290,4
AL Autauga County,01001,3080,5
AL Autauga County,01001,3157,6
AL Autauga County,01001,3330,7
```

Ignore first 0 rows ☐ First row is column name

Field Separator: ● Comma ○ Semicolon ○ Whitespace ○ Other

Group Indicator: ● Double Quotes ○ Single Quotes

The first screen shows the data. The file is comma separated and does not contain column names, so everything here is fine – hit Next.



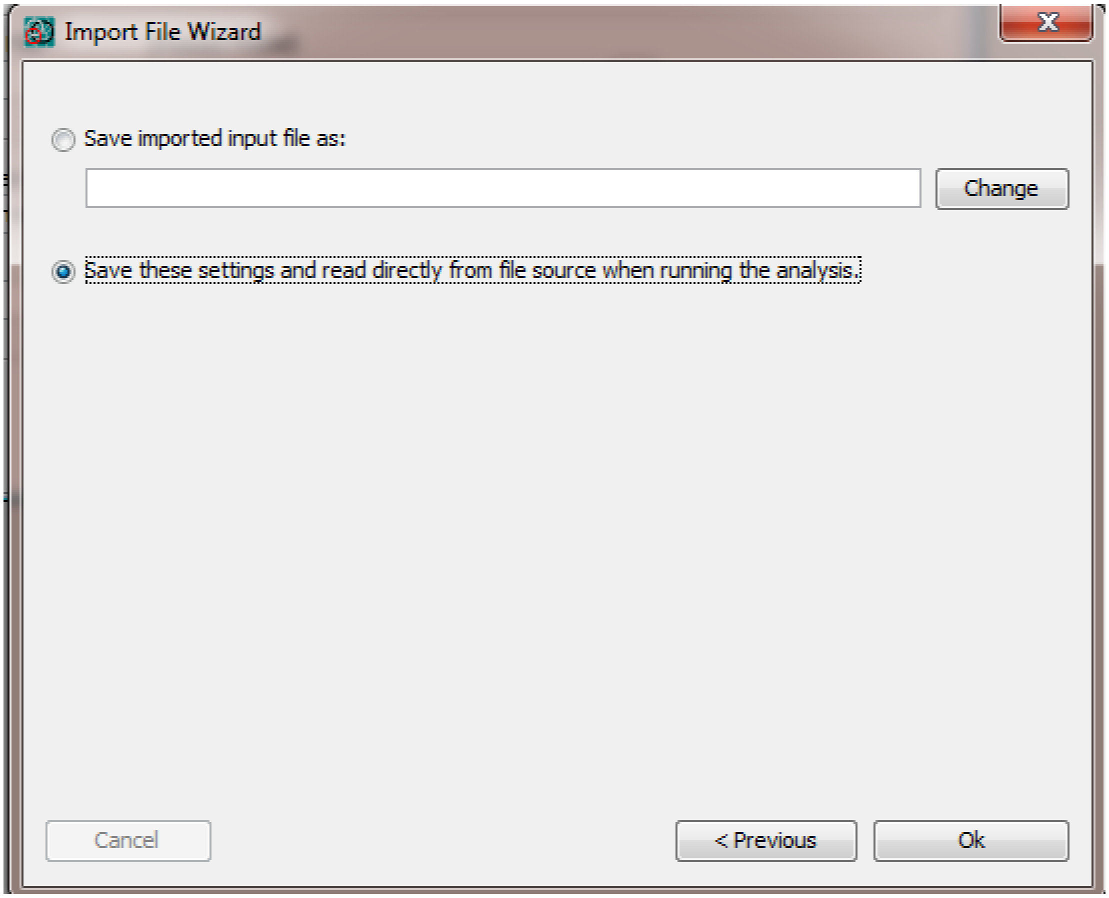
Display SaTScan Variables For: multinomial model

SaTScan Variable	Source File Variable
Location ID	Column 2
Number of Cases	Column 3
Date/Time (optional)	unassigned
Attribute (value)	Column 4

Generated Id #	One Count #	Column 1	Column 2	Column 3	Column 4
location1	1	AL Autauga County	01001	3579	1
location2	1	AL Autauga County	01001	3991	2
location3	1	AL Autauga County	01001	4290	3
location4	1	AL Autauga County	01001	4290	4
location5	1	AL Autauga County	01001	3080	5
location6	1	AL Autauga County	01001	3157	6
location7	1	AL Autauga County	01001	3330	7

= Column is not actually defined in file but can be used as SaTScan variable.

On the next screen we choose Column 2 (the five-digit state/county code) as the location ID, Column 3 (population) as the “number of cases”, and Column 4 (age group) as the Attribute. Hit Next again.



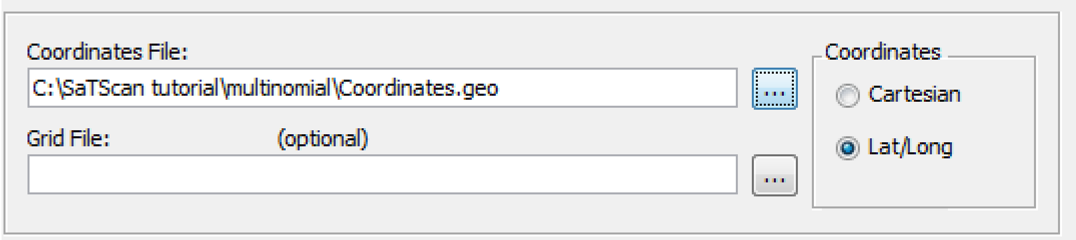
○ Save imported input file as:

● Save these settings and read directly from file source when running the analysis.

On the final screen, choose “save these settings and read directly from the file source when running the analysis”. Should you decide later to repeat the analysis with different parameters, this step will not have to be repeated.

6.2 Coordinates File (*.geo)

The process of specifying the coordinates file is nearly identical that of the case file.



Coordinates File: C:\SaTScan tutorial\multinomial\Coordinates.geo

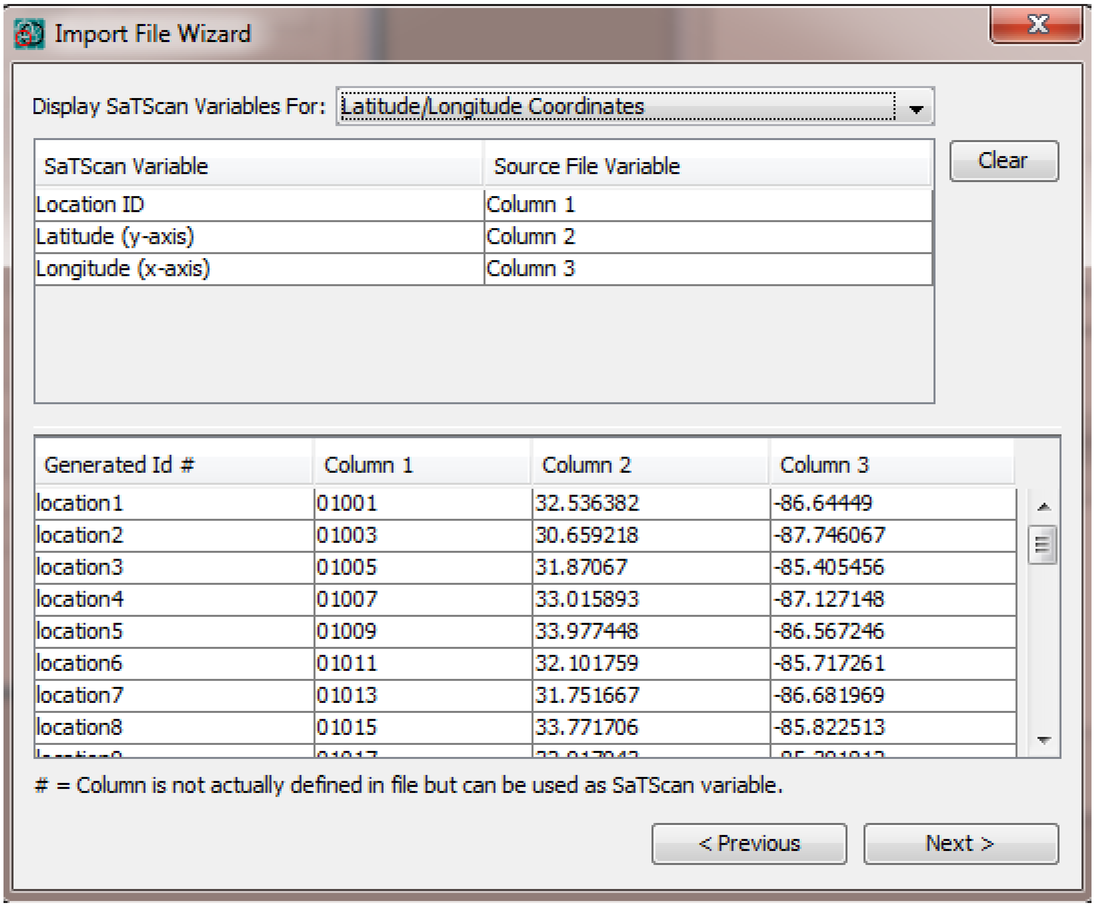
Grid File: (optional)

Coordinates
○ Cartesian
● Lat/Long

Click on the ... **button** to import the geographical coordinates data using the SaTScan file wizard.

Select coordinates.geo as the import file.

Then click Update and Next to get to this screen:



Display SaTScan Variables For: Latitude/Longitude Coordinates

SaTScan Variable	Source File Variable
Location ID	Column 1
Latitude (y-axis)	Column 2
Longitude (x-axis)	Column 3

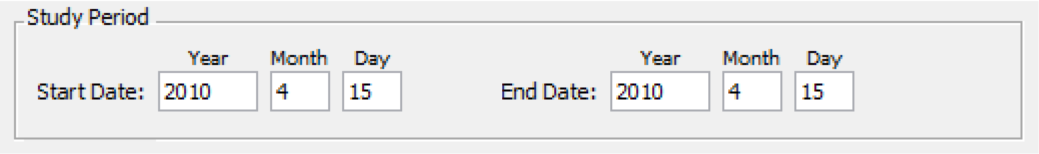
Generated Id #	Column 1	Column 2	Column 3
location1	01001	32.536382	-86.64449
location2	01003	30.659218	-87.746067
location3	01005	31.87067	-85.405456
location4	01007	33.015893	-87.127148
location5	01009	33.977448	-86.567246
location6	01011	32.101759	-85.717261
location7	01013	31.751667	-86.681969
location8	01015	33.771706	-85.822513

= Column is not actually defined in file but can be used as SaTScan variable.

Make sure that the Location ID is set to Column 1, latitude to Column 2, and longitude to Column 3 and hit Next. Again, choose “save these settings and read directly from the file source when running the analysis”.

6.3 Other information to specify on the Input tab

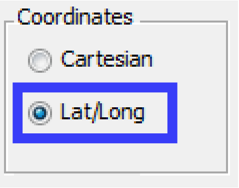
There are several other pieces of information to specify on the Input tab: Time Precision, Study Period, and Coordinates. Time Precision is only needed when a temporal analysis is being performed. Since we are executing a purely spatial analysis, no changes are needed: Time Precision can be left at the default setting of “none”. The study period is defined even for a purely spatial analysis, although we will not be making any use of it. Anything at all could be entered in here, but to be exact, we will enter the date the Census was taken:



Study Period

	Year	Month	Day		Year	Month	Day
Start Date:	2010	4	15	End Date:	2010	4	15

Finally, as the coordinates file contains latitude and longitude values, the button labeled Lat/Long should be selected:

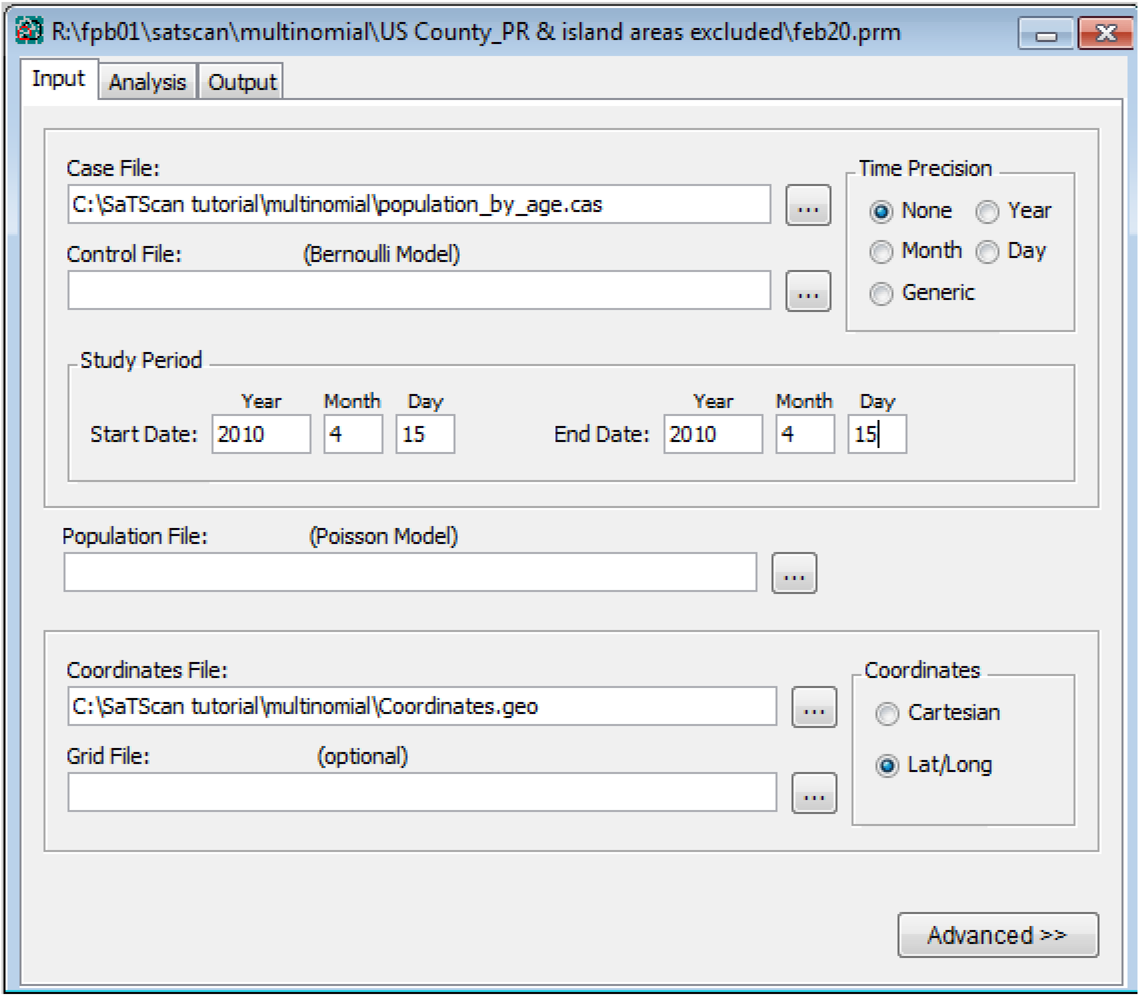


Coordinates

○ Cartesian

● Lat/Long

The final input screen should look like this:


R:\fpb01\satscan\multinomial\US County_PR & island areas excluded\feb20.prm

Case File: C:\SaTScan tutorial\multinomial\population_by_age.cas

Control File: (Bernoulli Model)

Time Precision: ● None ○ Year ○ Month ○ Day ○ Generic

Study Period

	Year	Month	Day
Start Date:	2010	4	15
End Date:	2010	4	15

Population File: (Poisson Model)

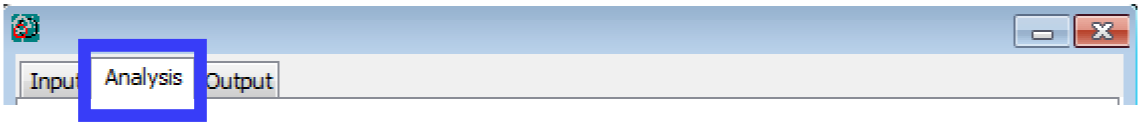
Coordinates File: C:\SaTScan tutorial\multinomial\Coordinates.geo

Grid File: (optional)

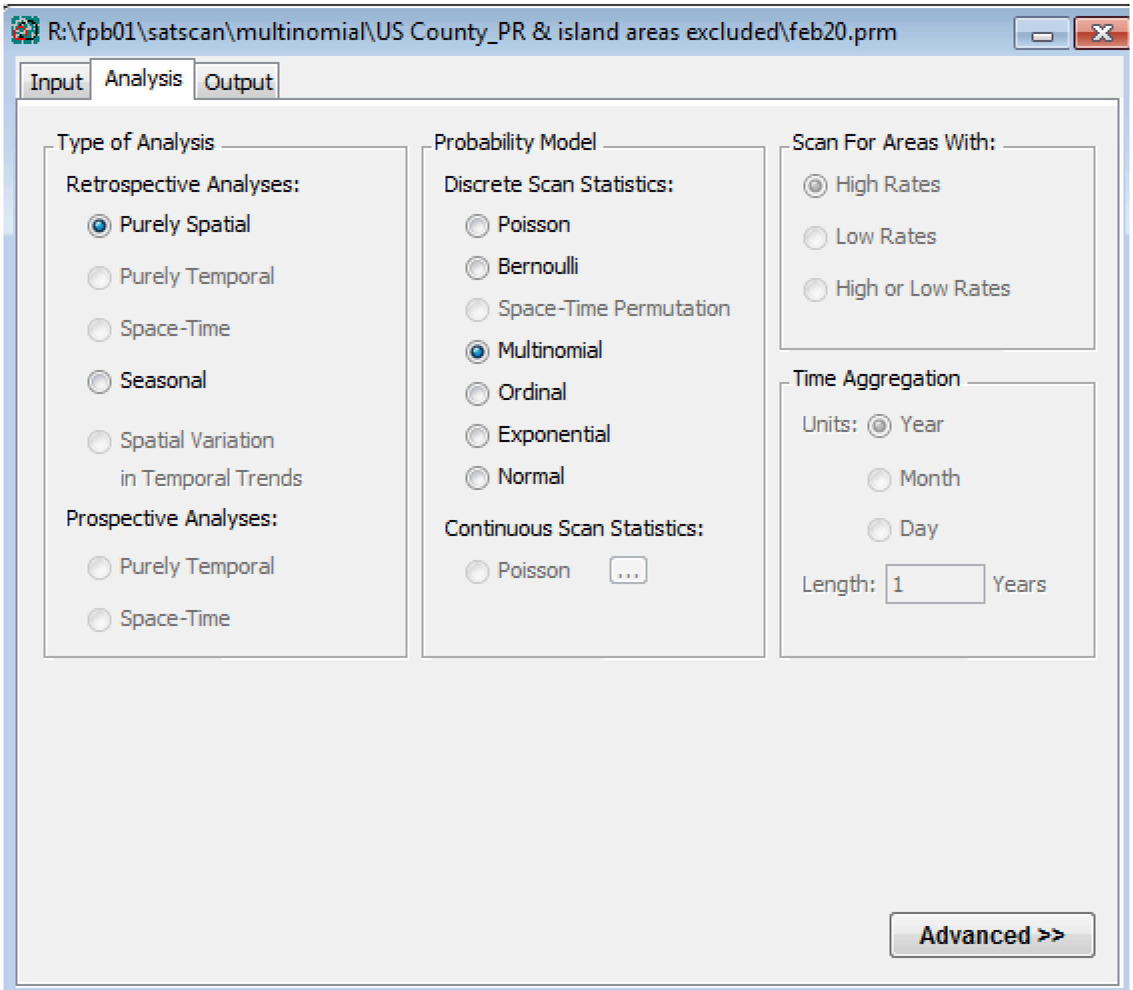
Coordinates: ○ Cartesian ● Lat/Long

7. Analysis Tab

Next we move to the analysis tab.



We choose a purely spatial analysis and a Multinomial probability model:



R:\fpb01\satscan\multinomial\US County_PR & island areas excluded\feb20.prm

Type of Analysis

Retrospective Analyses:

● Purely Spatial
○ Purely Temporal
○ Space-Time
○ Seasonal
○ 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:

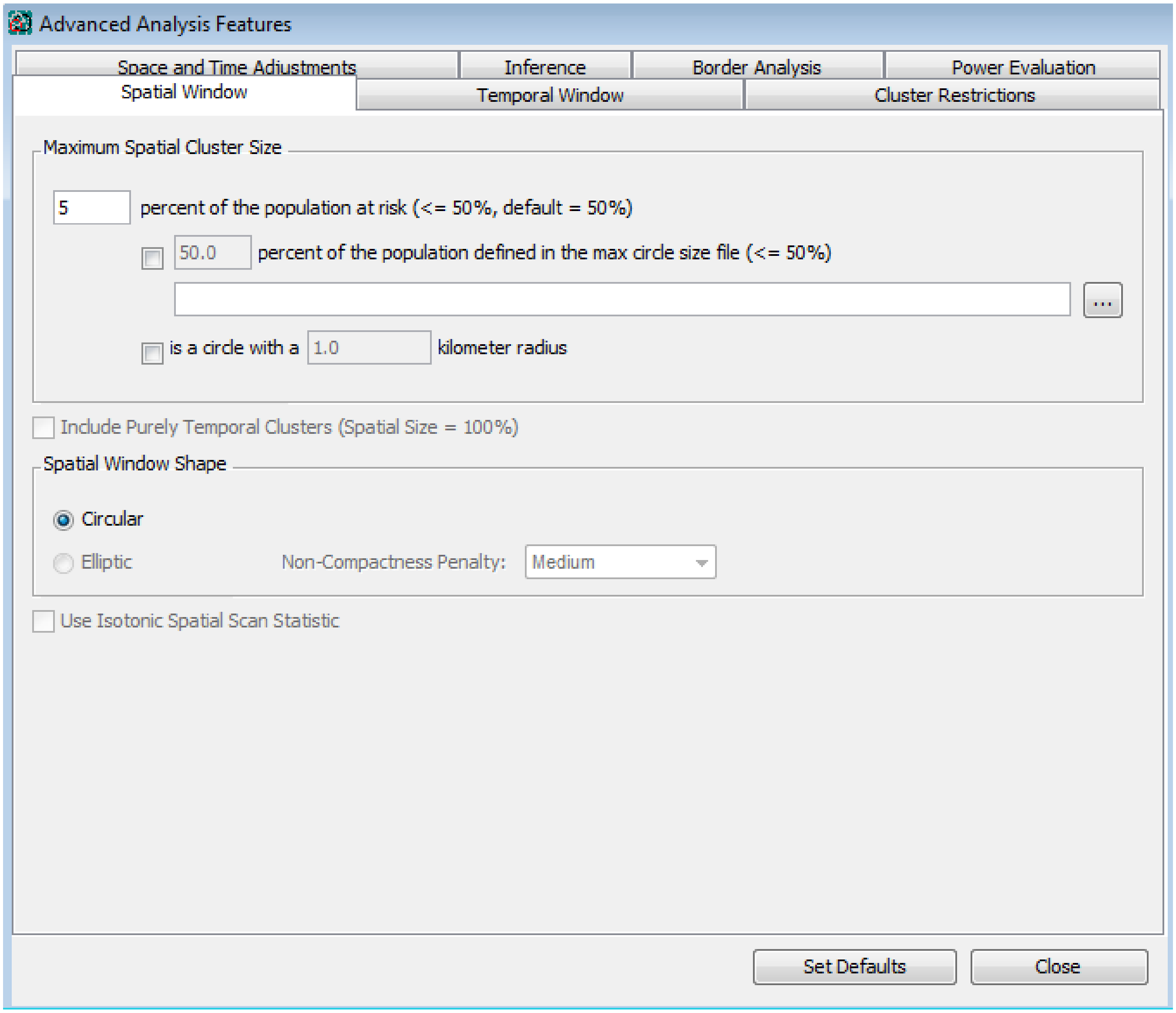
● High Rates
○ Low Rates
○ High or Low Rates

Time Aggregation

Units: ● Year
○ Month
○ Day

Length: 1 Years

The other sections of the Analysis tab do not apply to the Multinomial model and should be grayed out. Next click the ‘Advanced’ button on the lower right, which will open a window with seven tabs.



Maximum Spatial Cluster Size

5 percent of the population at risk (<= 50%, default = 50%)

□ 50.0 percent of the population defined in the max circle size file (<= 50%)

□ is a circle with a 1.0 kilometer radius

□ Include Purely Temporal Clusters (Spatial Size = 100%)

Spatial Window Shape

● Circular

○ Elliptic Non-Compactness Penalty: Medium

□ Use Isotonic Spatial Scan Statistic

Only two tabs are of interest. The first is the Spatial Window. Here, set the maximum spatial cluster size to 5 percent of the population at risk. This will limit the size of the resulting circles to a modest size. Note that with the default setting of 50%, half the United States could be considered a “cluster”. Using smaller numbers here tends to yield more interesting results, though there is no single “best” number to choose. Using numbers too small ends up excluding data for consideration – for example, Los Angeles County is about 3% of the U.S. population and could not be reported as part of any cluster if the maximum spatial cluster size was set below 3%.

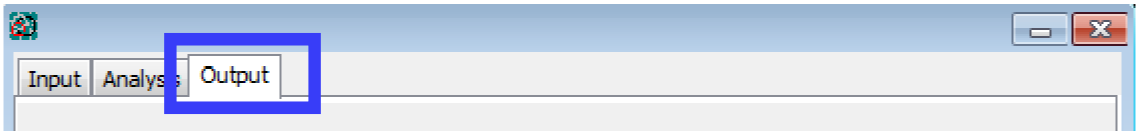
The second tab of interest is Inference, in particular the box which says Monte Carlo replications. The default number here is 999, which we will not change. This means that SaTScan will generate 999 random permutations of the data to compare with the real data. A certain degree of clustering can occur by chance. For real data to be identified as a significant cluster, it must be more unusual than at least 950 of the random permutations (this corresponds to a p-value of 50/1000, or 0.05).

The Multinomial model is computationally intensive, and 999 permutations of this data set could require several hours. (On the author's 2015-vintage Windows 7 desktop computer, it took about 90 minutes). We encourage you to run the analysis, work on other things, and check back in an hour or two. However, if this is not possible (for example, if it is being run within a classroom setting), the number of simulations can be changed to 9, which should require under a minute. This just means that precise p-values for the identified clusters will not be able to be calculated.

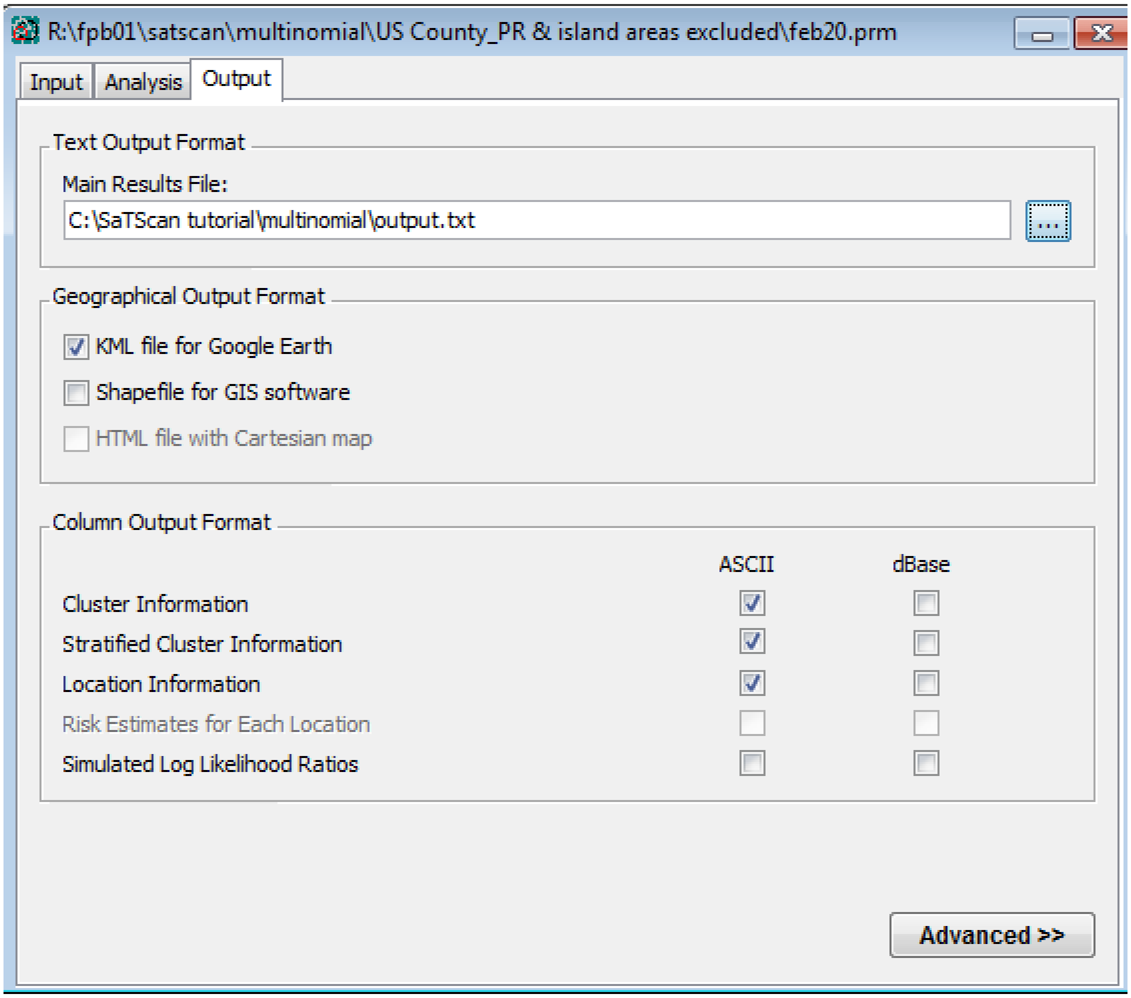
Use the Close button at the bottom left of the Advanced Analysis Features tabs.

8. Output Tab

SaTScan gives several options to view and save the results of the scan statistical analysis. You need to make these selections before you execute the SaTScan session. Click on the **Output** tab to see these options.



There are three sections on the main output tab corresponding to text, geographical, and column output formats. The main results file specified in the text output section consists of a detailed report that provides nearly all of the information of interest to a typical researcher. Specifically, it includes a summary of the data, location IDs of each location included in each cluster, the coordinates and radius of each cluster, and the population, number of cases, number of expected cases, relative risk and p-value for each cluster detected. In the geographical output section, you may choose to generate either a shapefile, readable by GIS software programs, or a KML file viewable in Google Earth. Finally, the column output section provides specific kinds of detailed information. These files are used less often but we recommend checking them anyhow, just in case there is a later need for them. ASCII refers to ordinary text file format while dBase is now a seldom-used format.



R:\fpb01\satscan\multinomial\US County_PR & island areas excluded\feb20.prm

Text Output Format

Main Results File: C:\SaTScan tutorial\multinomial\output.txt

Geographical Output Format

☑ KML file for Google Earth

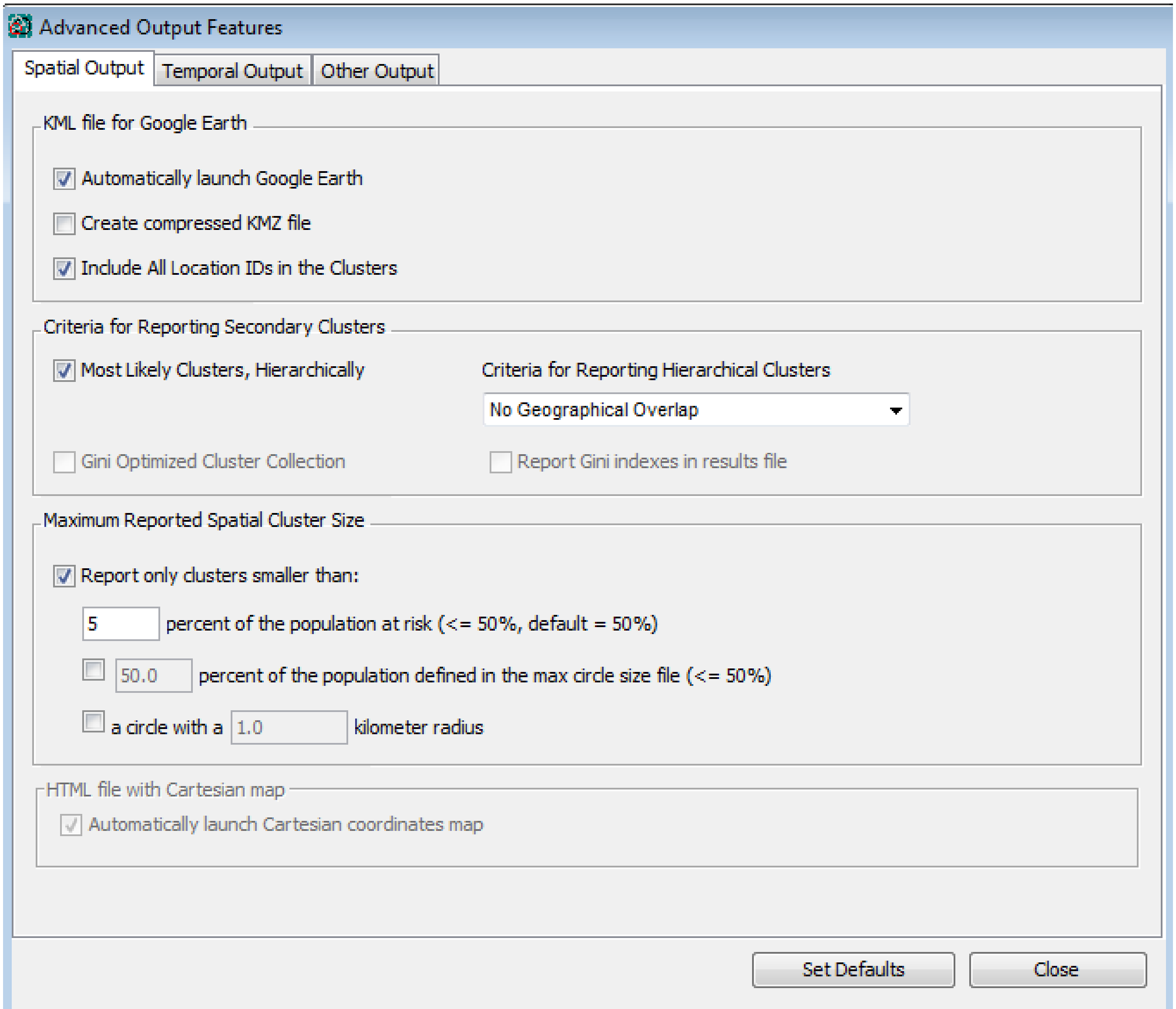
□ Shapefile for GIS software

□ HTML file with Cartesian map

Column Output Format

	ASCII	dBase
Cluster Information	☑	□
Stratified Cluster Information	☑	□
Location Information	☑	□
Risk Estimates for Each Location	□	□
Simulated Log Likelihood Ratios	□	□

SaTScan will typically find multiple overlapping clusters, most of which are nearly identical to each other. These can be filtered in the ‘Criteria for Reporting Secondary Clusters’ section of the Advanced Output Features. For this tutorial, select ‘No Geographical Overlap’, which is the most restrictive choice as well as the most common. With this option, a secondary cluster will only be reported if it does not overlap with a more likely cluster.



KML file for Google Earth

☑ Automatically launch Google Earth

□ Create compressed KMZ file

☑ Include All Location IDs in the Clusters

Criteria for Reporting Secondary Clusters

☑ Most Likely Clusters, Hierarchically

Criteria for Reporting Hierarchical Clusters: No Geographical Overlap

□ Gini Optimized Cluster Collection

□ Report Gini indexes in results file

Maximum Reported Spatial Cluster Size

☑ Report only clusters smaller than:

5 percent of the population at risk (<= 50%, default = 50%)

□ 50.0 percent of the population defined in the max circle size file (<= 50%)

□ a circle with a 1.0 kilometer radius

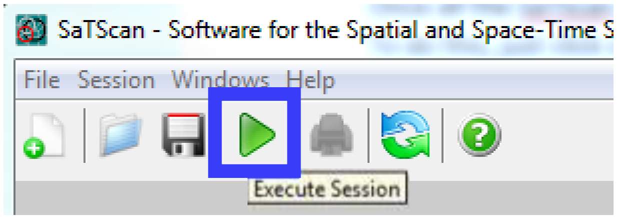
HTML file with Cartesian map

☑ Automatically launch Cartesian coordinates map

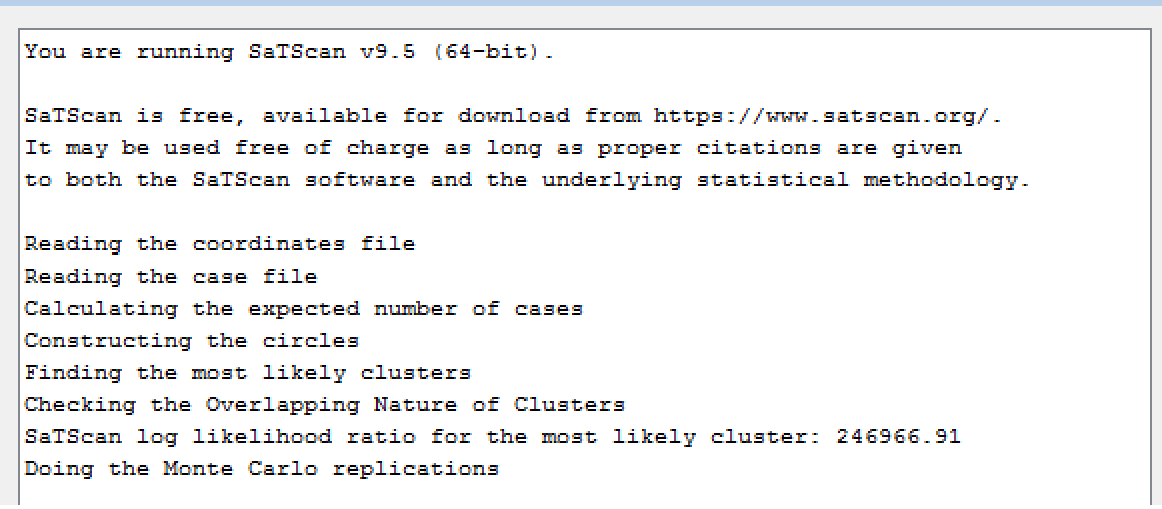
Also note the option here to automatically launch Google Earth. Check this if you have Google Earth installed on your computer. And now finally, it's time to run the analysis.

9. Executing SaTScan

To begin executing a SaTScan session, just click on the button with the green triangle.



A window will open which shows the progress being made:



```
You are running SaTScan v9.5 (64-bit).

SaTScan is free, available for download from https://www.satscan.org/.
It may be used free of charge as long as proper citations are given
to both the SaTScan software and the underlying statistical methodology.

Reading the coordinates file
Reading the case file
Calculating the expected number of cases
Constructing the circles
Finding the most likely clusters
Checking the Overlapping Nature of Clusters
SaTScan log likelihood ratio for the most likely cluster: 246966.91
Doing the Monte Carlo replications
```

Once the analysis is finished, this window will show the main results file, the same file specified in section 8. You will be able to scroll through this window to see all the results, as well as review the parameter settings you used.

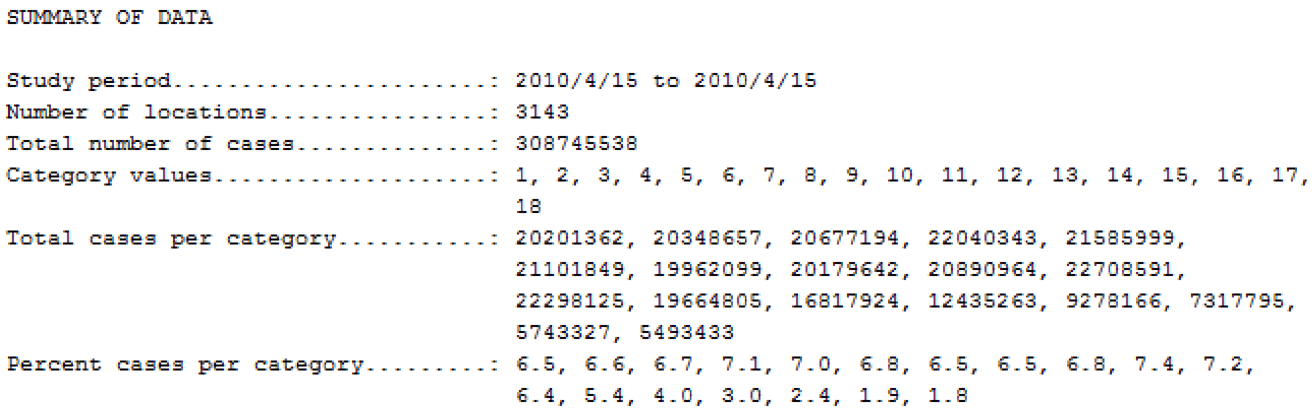
Sometimes SaTScan produces warning or error messages. The most common errors are problems with the input data, such as a location ID that is present in the case file but missing in the geographical coordinates file. The descriptions of the warning and errors are meant to help find problems that may exist in the input data. In this tutorial, you should not get any warnings or errors if you have done everything according to the tutorial instructions.

10. Interpreting the Results

10.1 Main results file

The main results file will automatically open once the execution is complete. In the top of the results file, it states the version and the time that SaTScan was run. It then should say: “Purely Spatial analysis scanning for clusters with all values using the Multinomial model.”

Next comes a summary of the data, which should look like this:

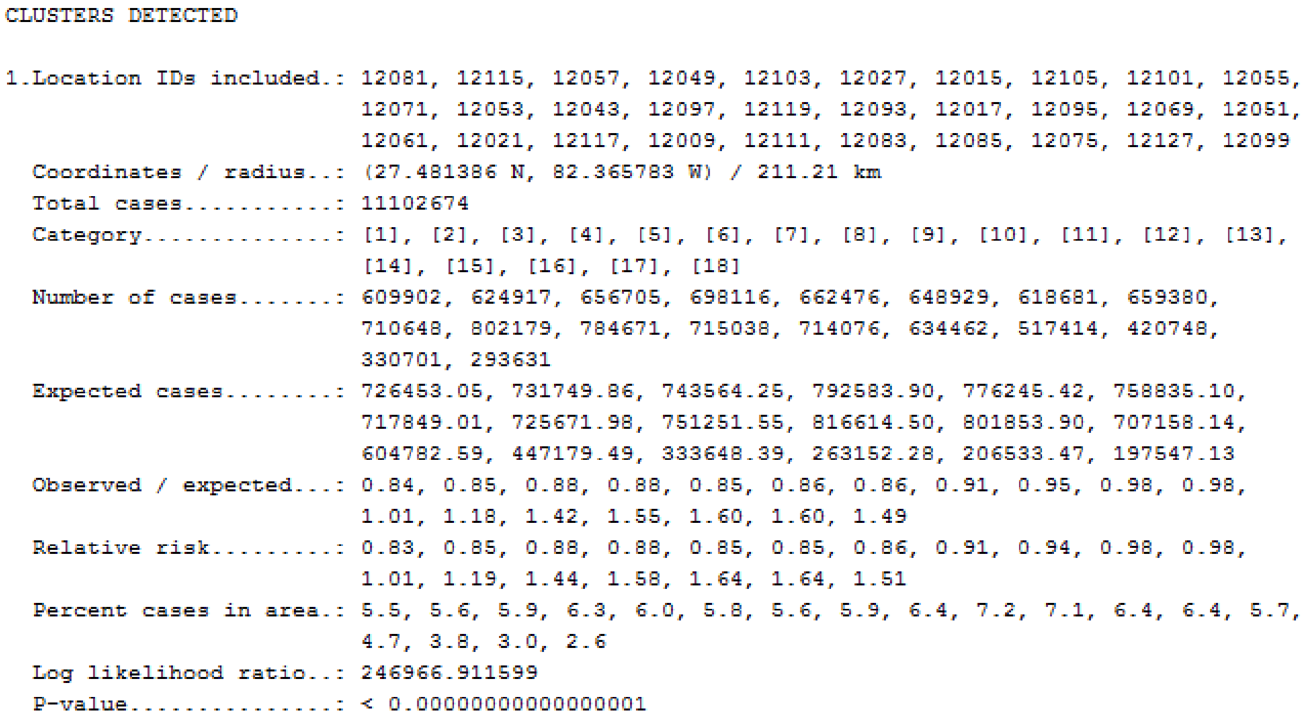


```
SUMMARY OF DATA

Study period.........................: 2010/4/15 to 2010/4/15
Number of locations..................: 3143
Total number of cases................: 308745538
Category values......................: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,
                                       18
Total cases per category.............: 20201362, 20348657, 20677194, 22040343, 21585999,
                                       21101849, 19962099, 20179642, 20890964, 22708591,
                                       22298125, 19664805, 16817924, 12435263, 9278166, 7317795,
                                       5743327, 5493433
Percent cases per category...........: 6.5, 6.6, 6.7, 7.1, 7.0, 6.8, 6.5, 6.5, 6.8, 7.4, 7.2,
                                       6.4, 5.4, 4.0, 3.0, 2.4, 1.9, 1.8
```

We see that there were 308.7 million total “cases” (population), 3143 locations (counties), and 18 category values (age groups). The percent cases per category shows that the oldest categories have the smallest populations, as would be expected.

The next section contains the cluster information. The most likely cluster (the cluster least likely to have occurred by chance) is given first, followed by the secondary clusters. For each cluster, a detailed description is provided. Here is the output for the first cluster:

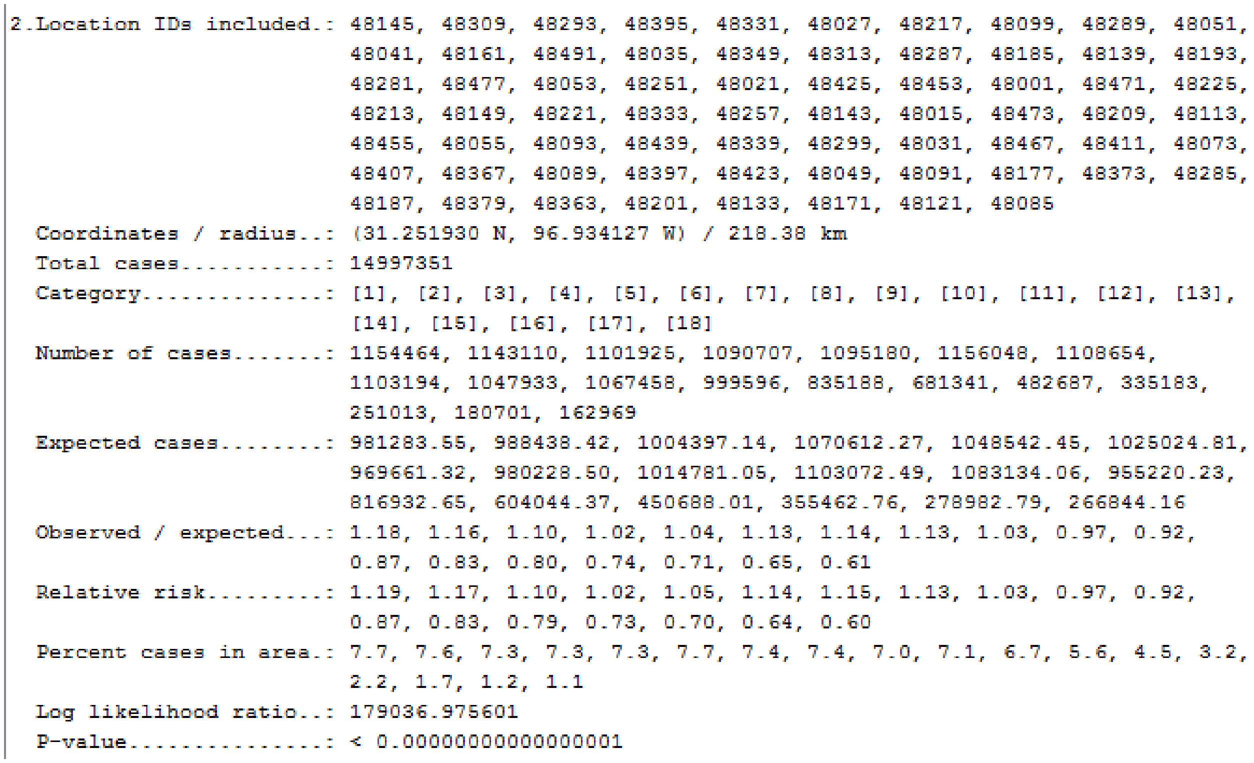


```
CLUSTERS DETECTED

1.Location IDs included.: 12081, 12115, 12057, 12049, 12103, 12027, 12015, 12105, 12101, 12055,
                          12071, 12053, 12043, 12097, 12119, 12093, 12017, 12095, 12069, 12051,
                          12061, 12021, 12117, 12009, 12111, 12083, 12085, 12075, 12127, 12099
  Coordinates / radius..: (27.481386 N, 82.365783 W) / 211.21 km
  Total cases...........: 11102674
  Category..............: [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13],
                          [14], [15], [16], [17], [18]
  Number of cases.......: 609902, 624917, 656705, 698116, 662476, 648929, 618681, 659380,
                          710648, 802179, 784671, 715038, 714076, 634462, 517414, 420748,
                          330701, 293631
  Expected cases........: 726453.05, 731749.86, 743564.25, 792583.90, 776245.42, 758835.10,
                          717849.01, 725671.98, 751251.55, 816614.50, 801853.90, 707158.14,
                          604782.59, 447179.49, 333648.39, 263152.28, 206533.47, 197547.13
  Observed / expected...: 0.84, 0.85, 0.88, 0.88, 0.85, 0.86, 0.86, 0.91, 0.95, 0.98, 0.98,
                          1.01, 1.18, 1.42, 1.55, 1.60, 1.60, 1.49
  Relative risk.........: 0.83, 0.85, 0.88, 0.88, 0.85, 0.85, 0.86, 0.91, 0.94, 0.98, 0.98,
                          1.01, 1.19, 1.44, 1.58, 1.64, 1.64, 1.51
  Percent cases in area.: 5.5, 5.6, 5.9, 6.3, 6.0, 5.8, 5.6, 5.9, 6.4, 7.2, 7.1, 6.4, 6.4, 5.7,
                          4.7, 3.8, 3.0, 2.6
  Log likelihood ratio..: 246966.911599
  P-value...............: < 0.00000000000000001
```

This cluster is centered at 27.48 north latitude and 82.37 west longitude and is 211 kilometers in diameter, covering much of central and southern Florida. The relative risks specified in the output indicate that this an area with an unusually high older population. There are 51% more people aged 85 and over relative to the rest of the nation, and 64% more in both the 75-79 and 80-84 groups. Meanwhile, there are 12-17% fewer children than would be expected. The p-value for the cluster is virtually zero – there is absolutely no way a result this extreme could occur by chance. This is consistent with what we know about Florida – it is a major retirement destination for the nation.

The number 2 cluster, in contrast, is skewed in the opposite direction:

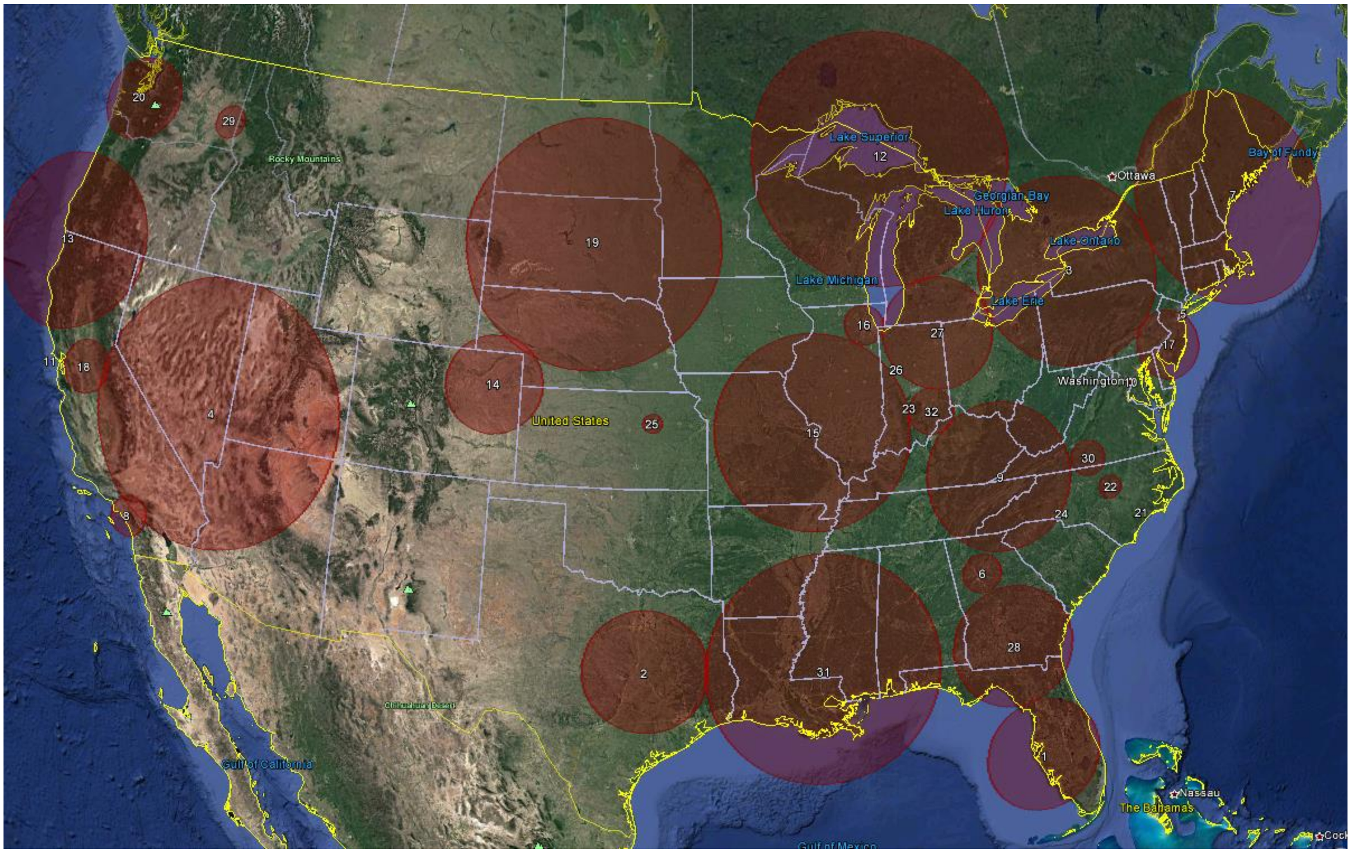


```
2.Location IDs included.: 48145, 48309, 48293, 48395, 48331, 48027, 48217, 48099, 48289, 48051,
                          48041, 48161, 48491, 48035, 48349, 48313, 48287, 48185, 48139, 48193,
                          48281, 48477, 48053, 48251, 48021, 48425, 48453, 48001, 48471, 48225,
                          48213, 48149, 48221, 48333, 48257, 48143, 48015, 48473, 48209, 48113,
                          48455, 48055, 48093, 48439, 48339, 48299, 48031, 48467, 48411, 48073,
                          48407, 48367, 48089, 48397, 48423, 48049, 48091, 48177, 48373, 48285,
                          48187, 48379, 48363, 48201, 48133, 48171, 48121, 48085
  Coordinates / radius..: (31.251930 N, 96.934127 W) / 218.38 km
  Total cases...........: 14997351
  Category..............: [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13],
                          [14], [15], [16], [17], [18]
  Number of cases.......: 1154464, 1143110, 1101925, 1090707, 1095180, 1156048, 1108654,
                          1103194, 1047933, 1067458, 999596, 835188, 681341, 482687, 335183,
                          251013, 180701, 162969
  Expected cases........: 981283.55, 988438.42, 1004397.14, 1070612.27, 1048542.45, 1025024.81,
                          969661.32, 980228.50, 1014781.05, 1103072.49, 1083134.06, 955220.23,
                          816932.65, 604044.37, 450688.01, 355462.76, 278982.79, 266844.16
  Observed / expected...: 1.18, 1.16, 1.10, 1.02, 1.04, 1.13, 1.14, 1.13, 1.03, 0.97, 0.92,
                          0.87, 0.83, 0.80, 0.74, 0.71, 0.65, 0.61
  Relative risk.........: 1.19, 1.17, 1.10, 1.02, 1.05, 1.14, 1.15, 1.13, 1.03, 0.97, 0.92,
                          0.87, 0.83, 0.79, 0.73, 0.70, 0.64, 0.60
  Percent cases in area.: 7.7, 7.6, 7.3, 7.3, 7.3, 7.7, 7.4, 7.4, 7.0, 7.1, 6.7, 5.6, 4.5, 3.2,
                          2.2, 1.7, 1.2, 1.1
  Log likelihood ratio..: 179036.975601
  P-value...............: < 0.00000000000000001
```

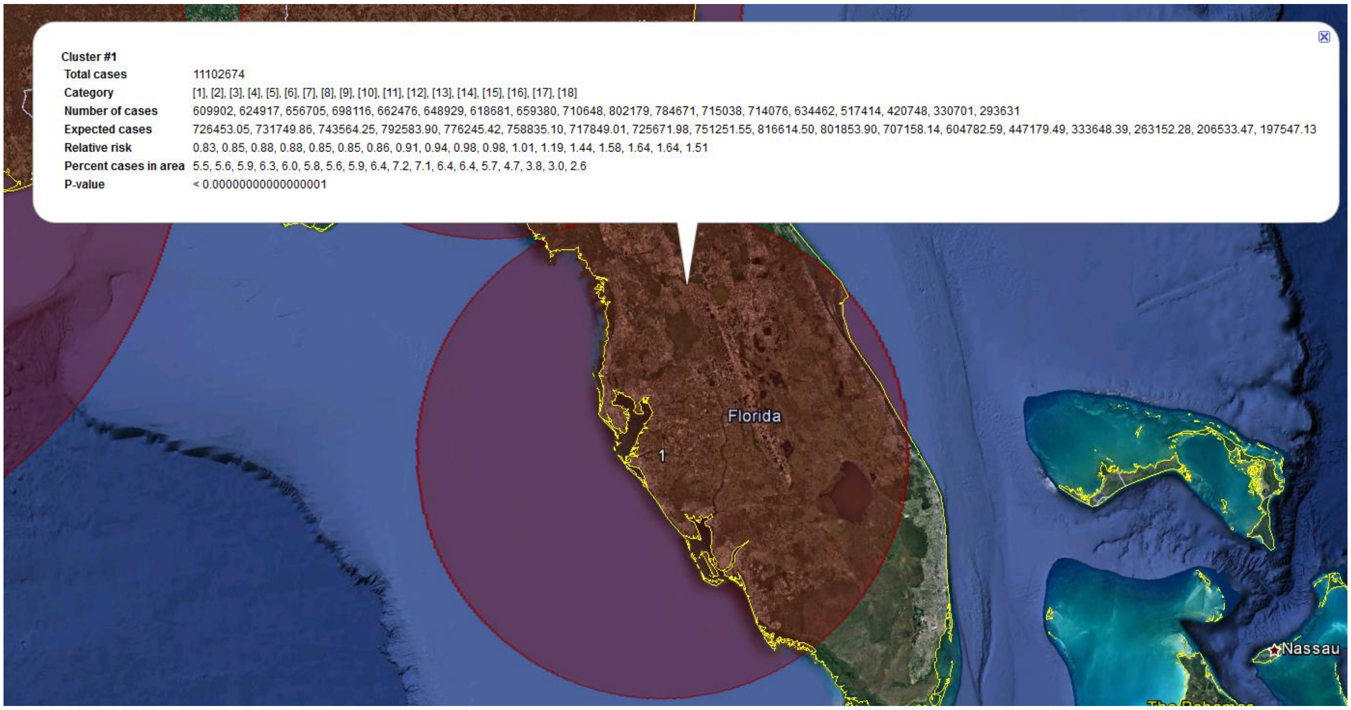
This cluster, in east-central Texas, identifies an area with disproportionately more children and fewer elderly than the nation as a whole. In all, you should see 32 cluster areas in total. Beneath the cluster output appears all of the parameter settings that were used, as well as the total run time.

10.2 Google Earth view

If you have Google Earth installed on your computer, SaTScan should have automatically opened it and presented the following map:



Clicking anywhere in a cluster opens a balloon containing all of the cluster information:



Cluster #1

Total cases	11102674
Category	[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]
Number of cases	609902, 624917, 656705, 698116, 662476, 648929, 618681, 659380, 710648, 802179, 784671, 715038, 714076, 634462, 517414, 420748, 330701, 293631
Expected cases	726453.05, 731749.86, 743564.25, 792583.90, 776245.42, 758835.10, 717849.01, 725671.98, 751251.55, 816614.50, 801853.90, 707158.14, 604782.59, 447179.49, 333648.39, 263152.28, 206533.47, 197547.13
Relative risk	0.83, 0.85, 0.88, 0.88, 0.85, 0.85, 0.86, 0.91, 0.94, 0.98, 0.98, 1.01, 1.19, 1.44, 1.58, 1.64, 1.64, 1.51
Percent cases in area	5.5, 5.6, 5.9, 6.3, 6.0, 5.8, 5.6, 5.9, 6.4, 7.2, 7.1, 6.4, 6.4, 5.7, 4.7, 3.8, 3.0, 2.6
P-value	< 0.00000000000000001

10.3 Supplemental output files

The additional output files generated by SaTScan mainly contain the same information as in the main results file, just formatted for convenience for importing into a spreadsheet or statistical software program. The cluster file (*.col.txt) arranges the cluster information in columns; the stratified cluster file (*.sci.txt) adds the statistics for each age group within each cluster; the location file (*.gis.txt) arranges the members of each cluster into columns, and the simulated log-likelihood ratio file (*.llr.txt) reports the test statistics obtained from the Monte Carlo simulations. Finally, the shape file (*.shp) can be imported into GIS software for viewing on a map.

11. Advantages of the multinomial model

You may be thinking that a simple binary model, say stratifying the population into over 65 and under 65, might have yielded the same result more simply. That is true of Florida, but not elsewhere. The multinomial model considers every possible combination of unusual age distribution, not simply old versus young. For example, cluster #5 consists of Manhattan, Brooklyn, Queens, and one county in northern New Jersey. This area has a relatively low proportion of children, an extremely high proportion of young adults aged 25-34 (30 to 40% more relative to the rest of the country), and typical numbers of senior citizens. Only the multinomial model is capable of identifying this distinctive pattern. The need to evaluate so many more combinations is the reason this method requires more time to run than the other statistical models.

12. References and Further Reading

This is the fifth in a series of SaTScan tutorials. We also recommend reviewing the first four, which describe the purely spatial Poisson model using cancer incidence data, the Bernoulli model using birth defects data, a second version of the Poisson tutorial that covers some of the advanced options within SaTScan, and the Ordinal model again using cancer data. These may be found at: https://www.satscan.org/tutorials.html

We also strongly recommend using the SaTScan User Guide. The User Guide is automatically downloaded together with the software, and can be found as a pdf file in the SaTScan directory. It can also be downloaded directly from the SaTScan website: https://www.satscan.org/techdoc.html.

For a more detailed description of the statistical theory behind the multinomial model as it is implemented within SaTScan, we recommend the paper “A Spatial Scan Statistic for Multinomial Data” by Inkyung Jung, Martin Kulldorff and Otukei John Richard, published in the journal *Statistics in Medicine* in 2010 (volume 29, pages 1910-1918). Please note that this paper is much more technical than the language used in this tutorial.