

# Ordinal Scan Statistic for Identifying Unusual Cancer Stage Patterns

Francis Boscoe, New York State Department of Health

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

July, 2017

## 1. Introduction

SaTScan™ 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 ordered categorical data. The tutorial can be utilized independently, without first having worked through other SaTScan tutorials. It was written using SaTScan version 9.4.4 for Windows, released in August, 2016. 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 of colorectal cancer diagnosis in New York State, USA, in order to determine if there are any geographical clusters of either earlier or more advanced stage at diagnosis. That is, we will determine if there are any geographical areas where the distribution of cancer stage is unusually skewed relative to the statewide average.

## 2. What is the Ordinal 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 ordered categories, known as the ordinal model. This method considers all possible ways that the categories can be grouped while preserving their order, and thus can provide more information than a purely binary analysis. For example, we might be interested in identifying regions where the population tends to skew older and might require additional social services. This is often done by dividing the population into those over and under some specific age like 55 or 60 or 65. However, there are major differences between these ages and people who are 75 and 85 and 95, who require increasing levels of services, so it might be more informative to consider all the 5-year age groups within the same model at the same time.

### **3. New York State Colorectal Cancer Data**

Colorectal cancer is a cancer in which malignant cells form in the colon and rectum. Colorectal cancer is presently the 4th most common incident cancer (after breast, prostate, and lung) and 2nd leading cause of cancer death (after lung) in New York State. After cancer has been diagnosed, tests are done to find the extent to which cancer cells have spread within the colon and rectum or to other parts of the body, which is called staging. Cancer stage is one of important prognostic factors to predict cancer patients' chance of survival.

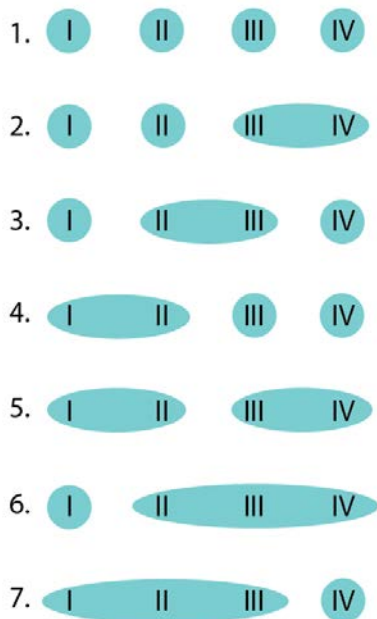
Colorectal cancer is one of the cancer types most amenable to screening. Screening through fecal occult blood test (FOBT), colonoscopy, or sigmoidoscopy at prescribed intervals beginning at age 50 has been proven to reduce both incidence and mortality by detecting tumors earlier, including before they even become invasive. This means that more screening in the future could help reduce both the frequency and severity of cancer diagnosis. For this reason, it is useful to analyze the stage distribution, as areas where there is more screening should have a more favorable stage distributions, other things being equal. Since stage can be represented by ordered categories, it is a good application of the ordinal model with SaTScan.

For this tutorial, tumors have been categorized into four stages – stages I, II, III, and IV, where increasing numbers represent more severe disease. This staging system is known as the American Joint Committee on Cancer (AJCC) staging system and is one of several staging systems currently in wide use, and lends itself to an analysis using the ordinal model. The stages are defined on the American Cancer website here:

<https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.html>

Following this link, you will notice that aside from stage I, each stage contains several substages. These have been excluded from the tutorial data set in the interest of simplicity.

The ordinal model will be able to test not only where there are unusual patterns of stage IV cancers (the usual application of SaTScan), but seven different possible relationships, as illustrated below:



For example, the fifth row tests for whether there are areas where the proportions of stages I and II combined are unusual relative to the proportions of stages III and IV combined.

The data used in this tutorial consist of 43,972 colorectal tumors diagnosed among New York State residents between 2010 and 2014 and reported to the New York State Cancer Registry, which is a part of the New York State Department of Health. The number of tumors by stage are as follows:

Stage I: 10,910 (24.9%)

Stage II: 11,920 (27.2%)

Stage III: 12,147 (27.7%)

Stage IV: 8,815 (20.1%)

The tumors have been grouped geographically into 952 groups of 46 tumors each. Each group has a single point location, which is the average location of the tumors in the group. The tumors in a group do not necessarily belong to the same county, city, zip code or any other geographic unit. This grouping method was chosen to protect patient confidentiality and to allow execution of SaTScan in a reasonable amount of time. Of the various statistical models available in SaTScan, the ordinal model tends to be more computationally intensive.

*Input Data File:* The colorectal cancer data file for this tutorial can be downloaded from the SaTScan website <https://www.satscan.org/datasets/nyscolorectal/index.html>. It is a text file, with the name colorectal\_cancer.csv (a comma-separated text file). In addition to being used in SaTScan, it can be opened in Excel or any text editor. The file contains five columns: case ID number (1-43972), group number (1-952), latitude, longitude, and stage.

*Data directory structure:* The tutorial requires creating several folders, as follows:

Main Directory: C: \SaTScan tutorial\**Ordinal Model**

SaTScan Input Files: C: \SaTScan tutorial\Ordinal Model\**InputFiles**

SaTScan Output Files: C: \SaTScan tutorial\Ordinal Model\**OutputFiles**

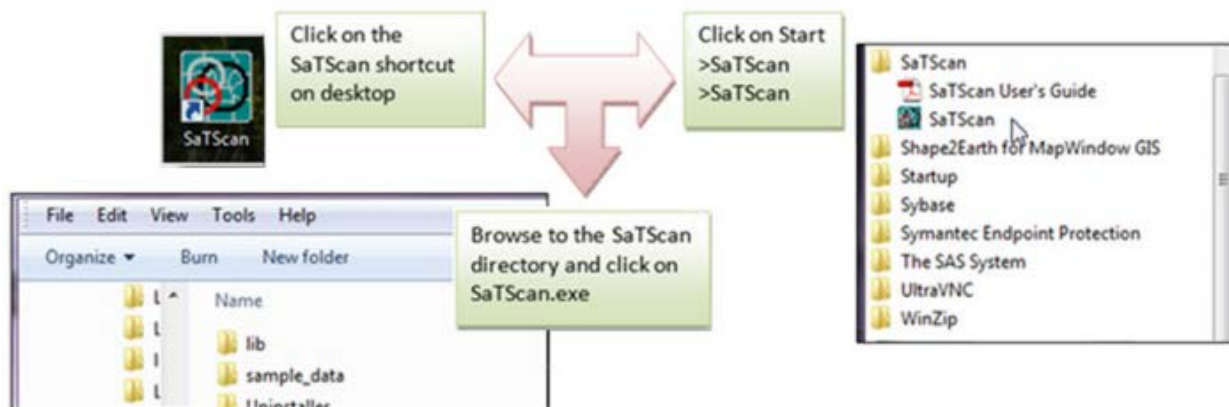
#### 4. SaTScan Software Download and Installation

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.

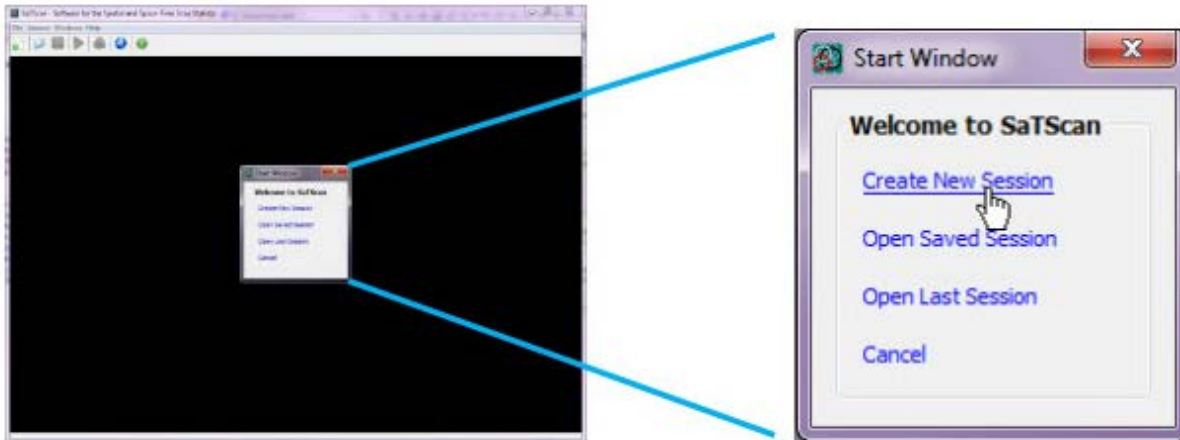
The SaTScan software is able to depict clusters with Google Earth. This requires Google Earth to be installed on your computer. To download, go to <https://www.google.com/earth/index.html>, and follow the instructions. This is an optional step; SaTScan can be run without Google Earth.

#### 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.

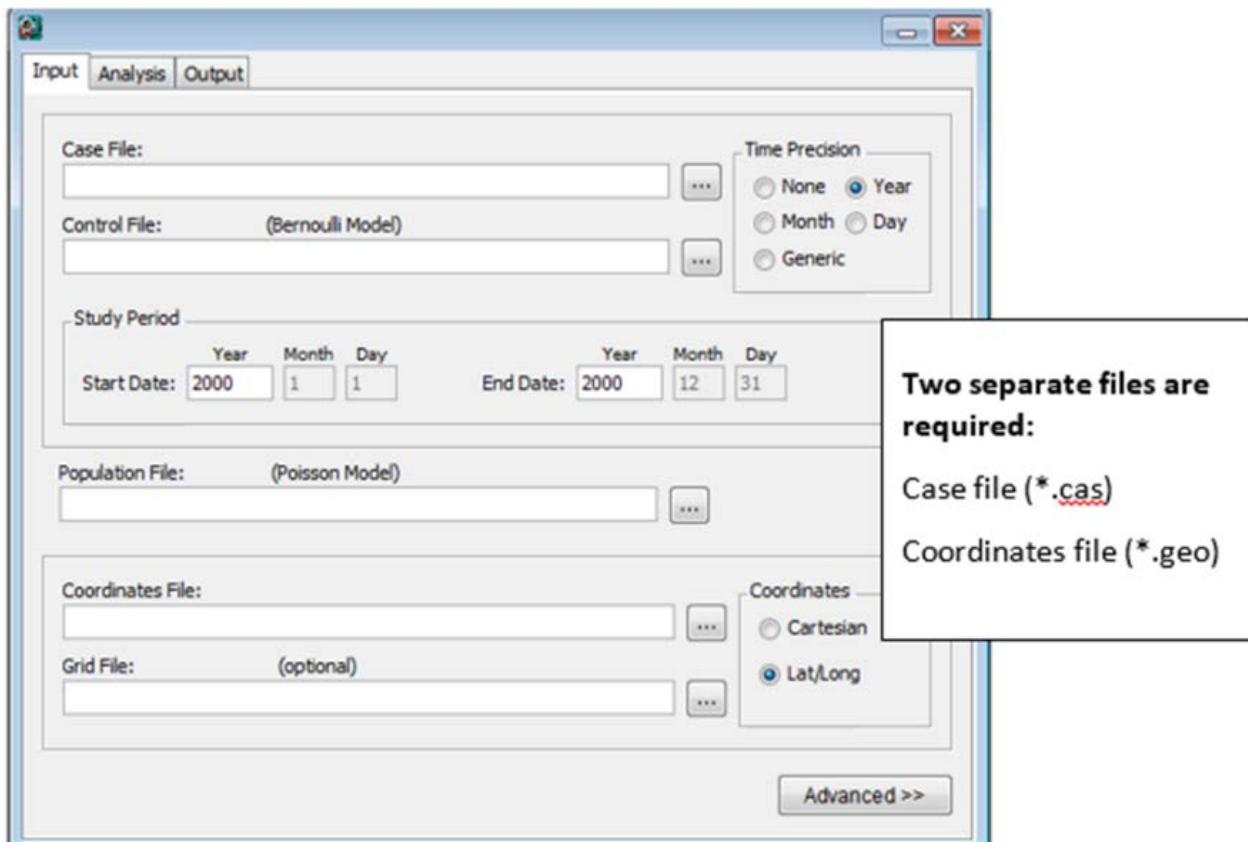


Next, select “Create New Session” from the Start Window.



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.

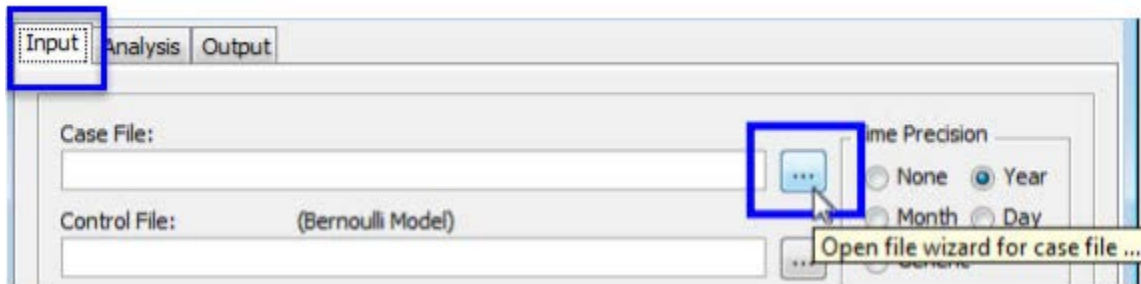
## 6. Input Tab




The first of the three main tabs is used to specify the input data. For ordered categorical data with the Ordinal model, two input files are required: a case file containing information about the cancer cases, and a coordinates file containing the latitude and longitude coordinates. These two files will be created from the colorectal\_cancer.csv file.

## 6.1 Case File (\*.cas)

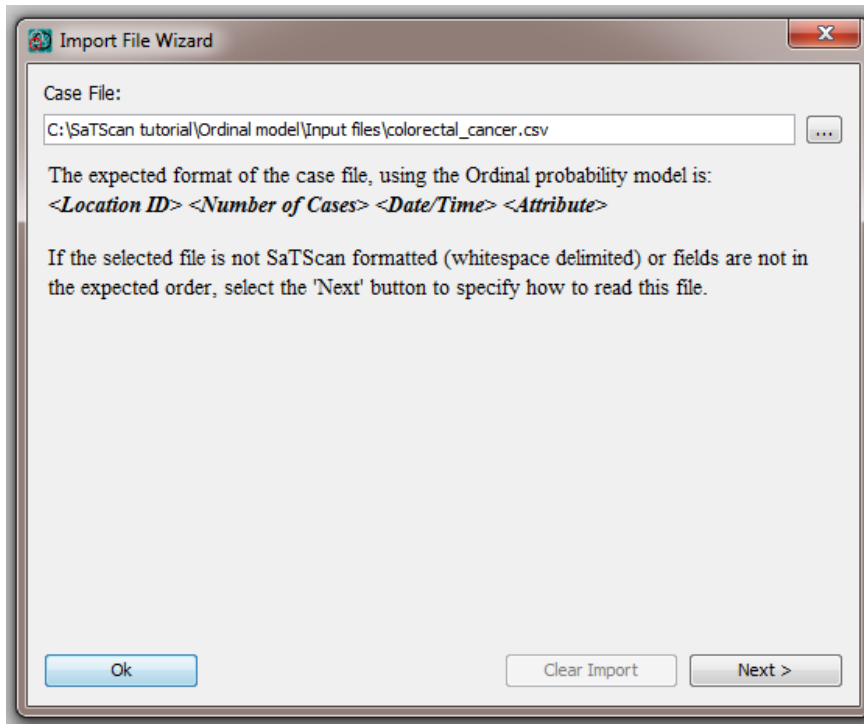
If the Case File is already in the SaTScan input file format (see User Guide), it is enough to specify its name in the Case File text box. In our case, this is not true, so we will use the SaTScan Import Wizard.



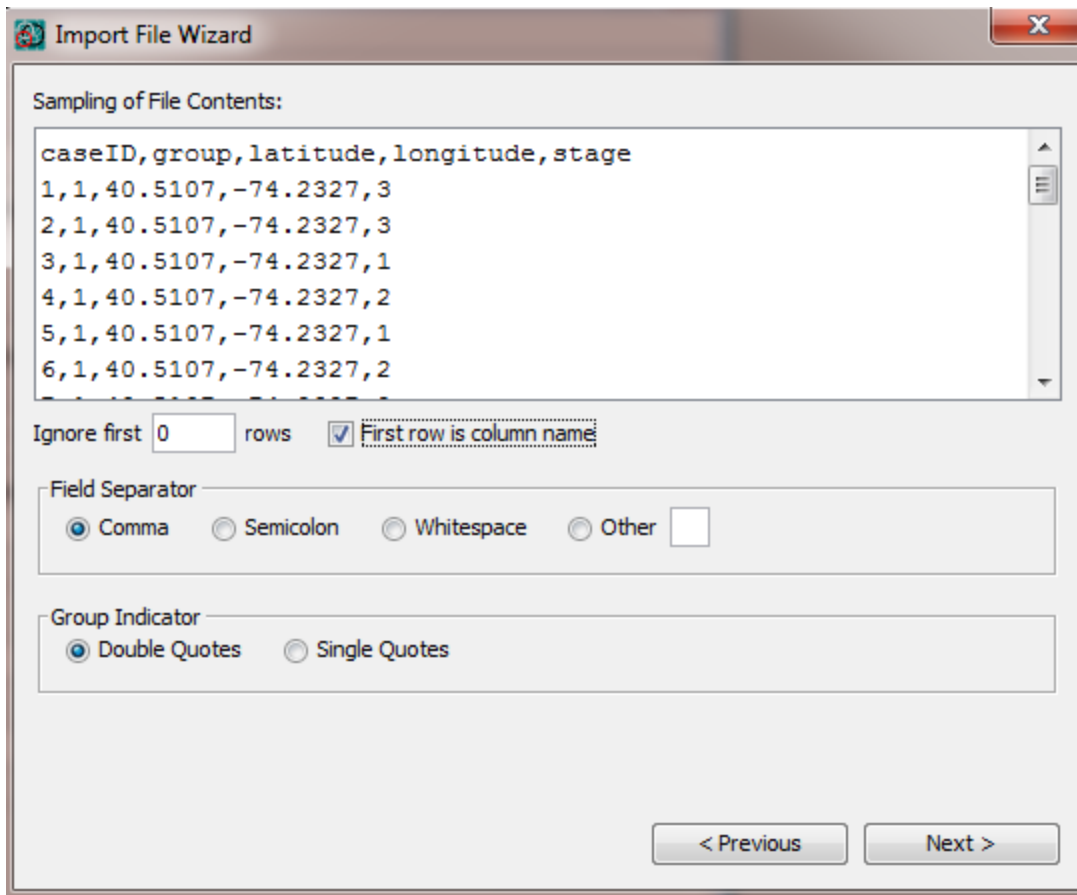
Click on the  button to import the case data using the SaTScan file wizard.

Select colorectal\_cancer.csv file as the import file.

The SaTScan import wizard can read several common file formats including \*.csv, \*.xlsx, \*.xls, \*.dbf, \*.txt and \*.shp. We are importing from a \*.csv file. Navigate to the colorectal\_cancer.csv file and open it, which will bring you to the Import File Wizard screen:

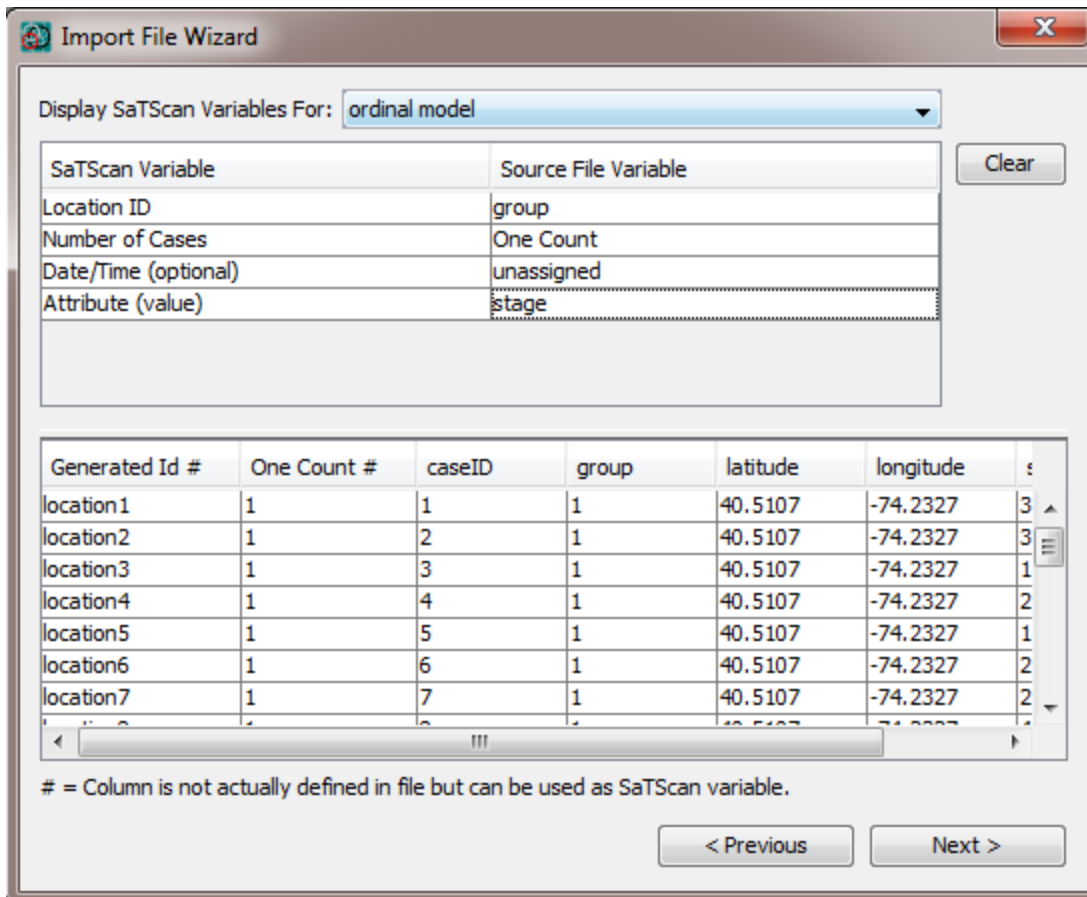


Click **Next** for a view of the file contents. Check the box which says “first row is column name” and choose comma as the field separator, then click Next again.

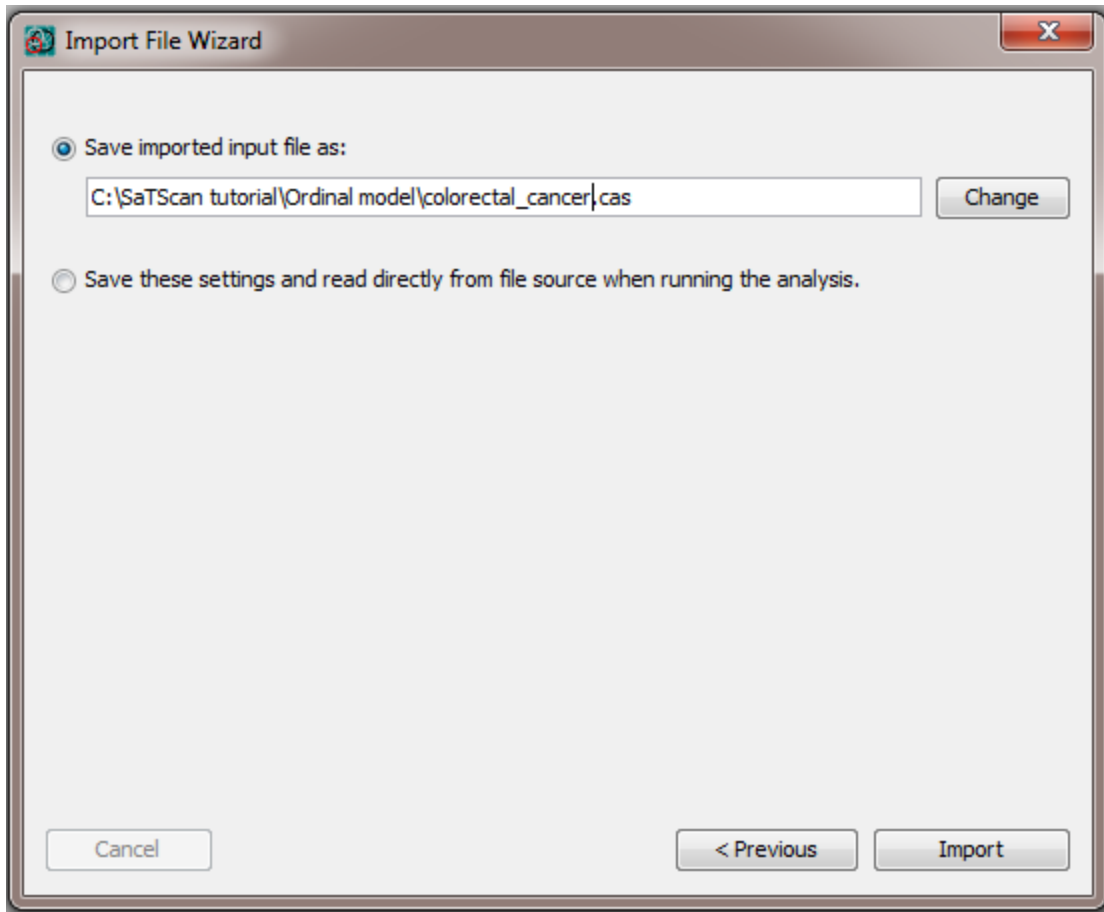


On the next screen, in the topmost box, choose Ordinal Model. Then choose the four variables needed for the Case file. Choose “group” for the location ID (representing the 952 grouped locations), “One Count” for the number of cases (meaning that each row is an individual case), and “stage” for the Attribute. We are not performing a temporal analysis, so leave Date/Time unassigned. Click **Next** again.





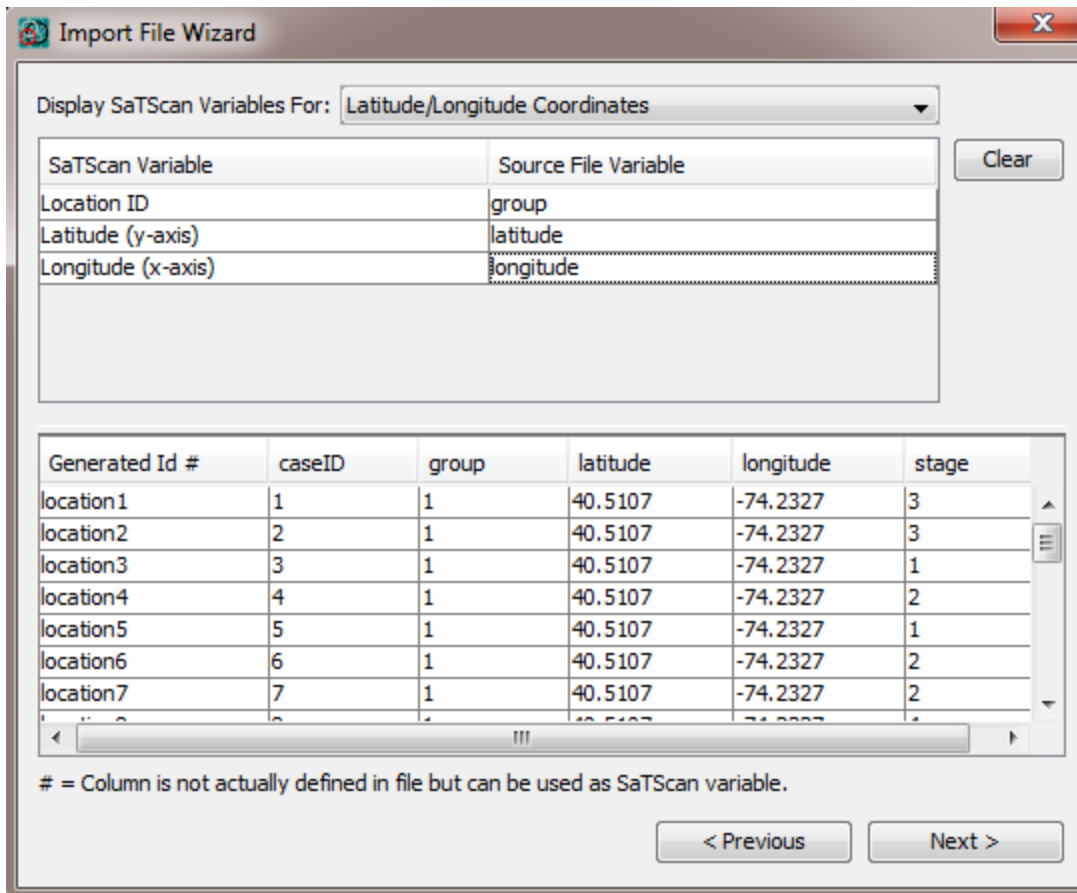
On the last screen of the wizard, you have the option of saving the file as a .cas file, or allowing the software to read directly from the source file when executing the analysis. We will choose to save the file, to save time by not needing to repeat the Import Wizard steps if we later need to repeat the analysis:



Finally, click Import.

## 6.2 Geographical Coordinates File (\*.geo)

A very similar process is used to create the coordinates file. The coordinates in the input file are given in latitude and longitude, which is the default in SaTScan. Repeat the steps above until reaching the screen where you choose the variables needed for the coordinates file. Again, “group” is the Location ID, and latitude and longitude are the variables of the same names:



Again, save the file and click Import.

### 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”.

#### 6.3.1 Study Period

Study period is defined even for a purely spatial analysis, so that disease rates can be properly calculated. In the colorectal\_cancer.csv file, the time period is January 1, 2010 through December 31, 2014:

Study Period

	Year	Month	Day		Year	Month	Day
Start Date:	2010	1	1	End Date:	2014	12	31

### 6.3.2 Coordinates

In SaTScan, the geographical locations can either be specified as latitude and longitude, or as Cartesian coordinates, such as when the data are projected using a Universal Transverse Mercator (UTM) projection. Our data set uses latitude and longitude coordinates. Even though this was specified during the import process, it must also be specified again on the input tab. Since latitude and longitude is the default, it should already be selected:

Coordinates

Cartesian

Lat/Long

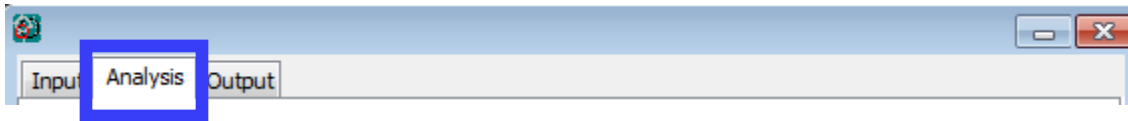
We are now finished with the input tab, which should look like this:

The screenshot shows the SaTScan software interface with the 'Input' tab selected. The configuration is as follows:

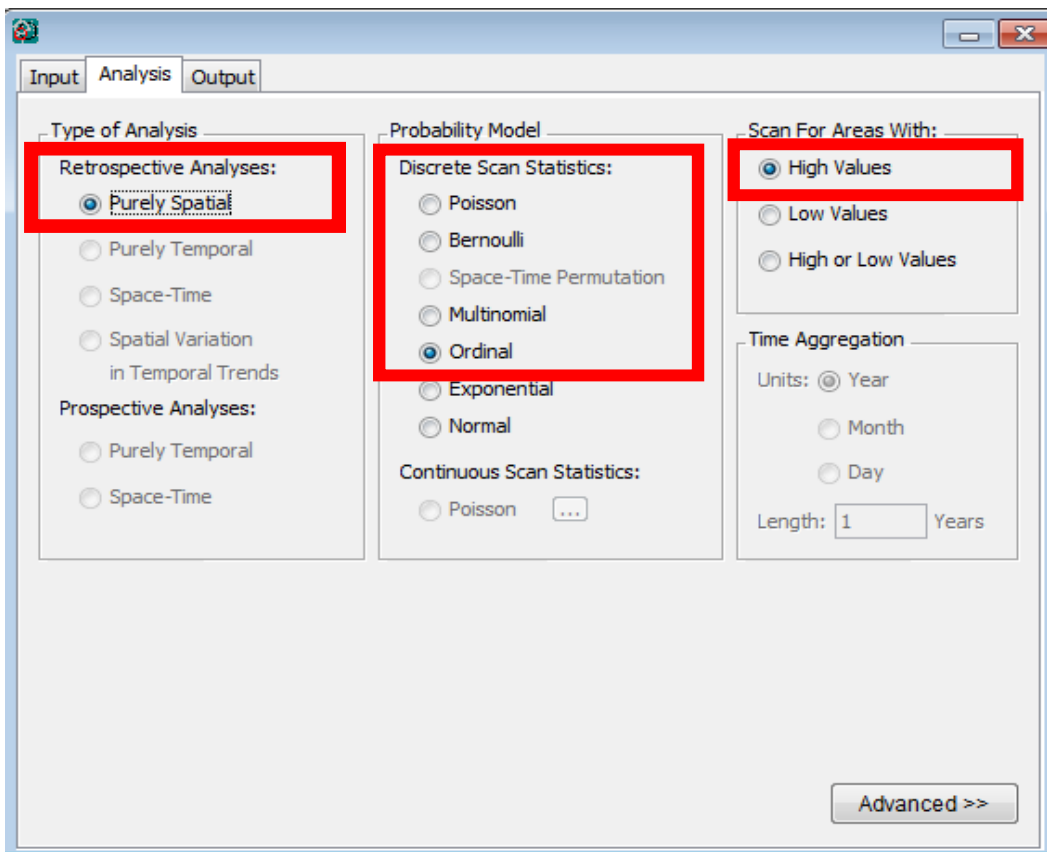
- Case File:** C:\SaTScan tutorial\Ordinal model\colorectal\_cancer.cas
- Control File:** (Bernoulli Model)
- Time Precision:** None (selected), Year, Month, Day, Generic
- Study Period:** Start Date: 2010-01-01, End Date: 2014-12-31
- Population File:** (Poisson Model)
- Coordinates File:** C:\SaTScan tutorial\Ordinal model\colorectal\_cancer.geo
- Grid File:** (optional)
- Coordinates:** Lat/Long (selected), Cartesian
- Advanced >>** button is visible at the bottom right.

## 7. Analysis Tab

Next we move to the analysis tab.

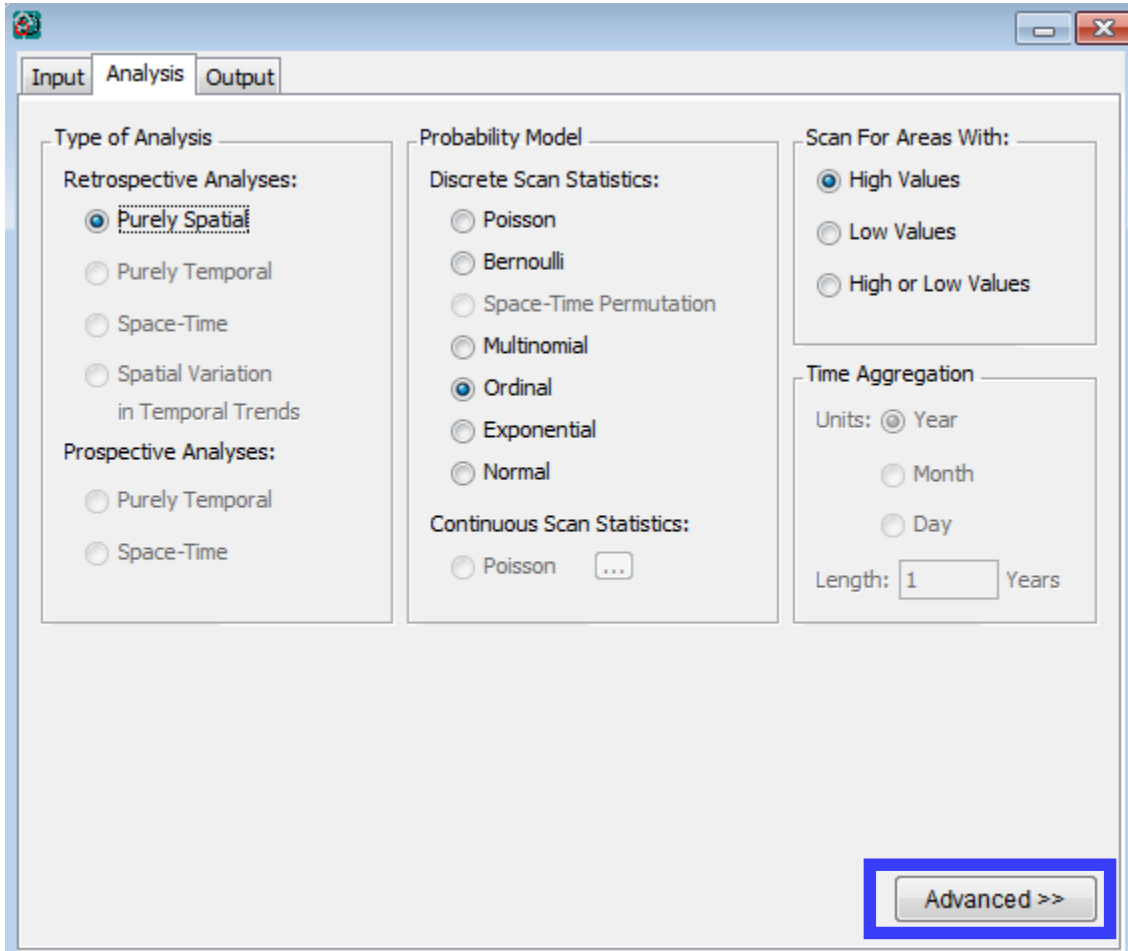


We choose a purely spatial analysis and an Ordinal probability model. We then have the option of scanning for area with high values (a stage distribution skewed toward more advanced stages), low values (a stage distribution skewed toward earlier stages) or both. Since we are primarily interested in areas with more advanced stages, we choose 'high values'.

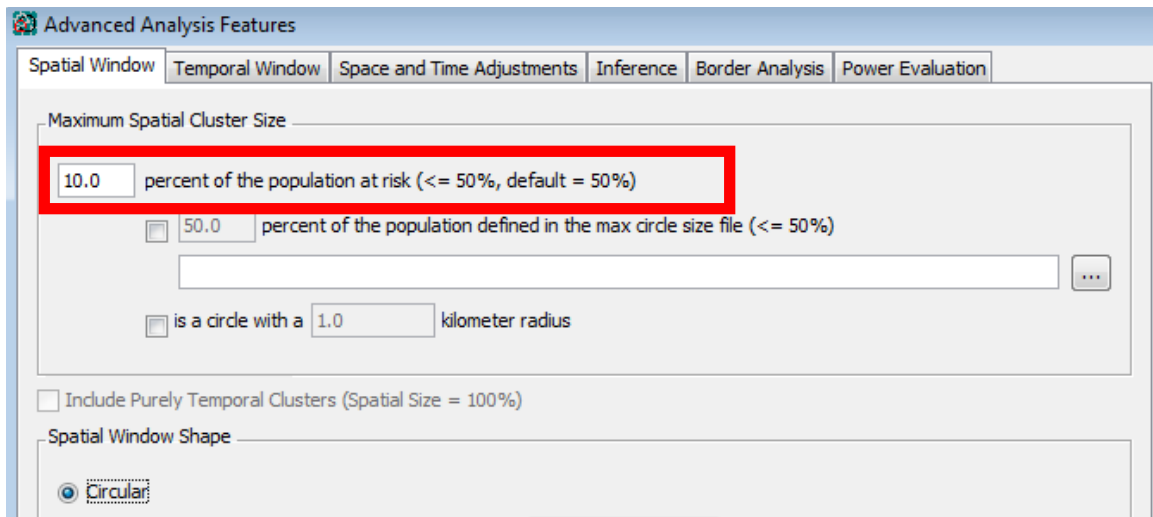


The last box on the Analysis Tab is for Time Aggregation, but that is only relevant for purely temporal and space-time analyses. Since we are conducted a purely spatial analysis, this option is greyed out, and can be ignored.

Before moving to the Output Tab, we will take a look at the advanced analysis features, by clicking on 'Advanced' in the bottom right corner of the Analysis Tab.

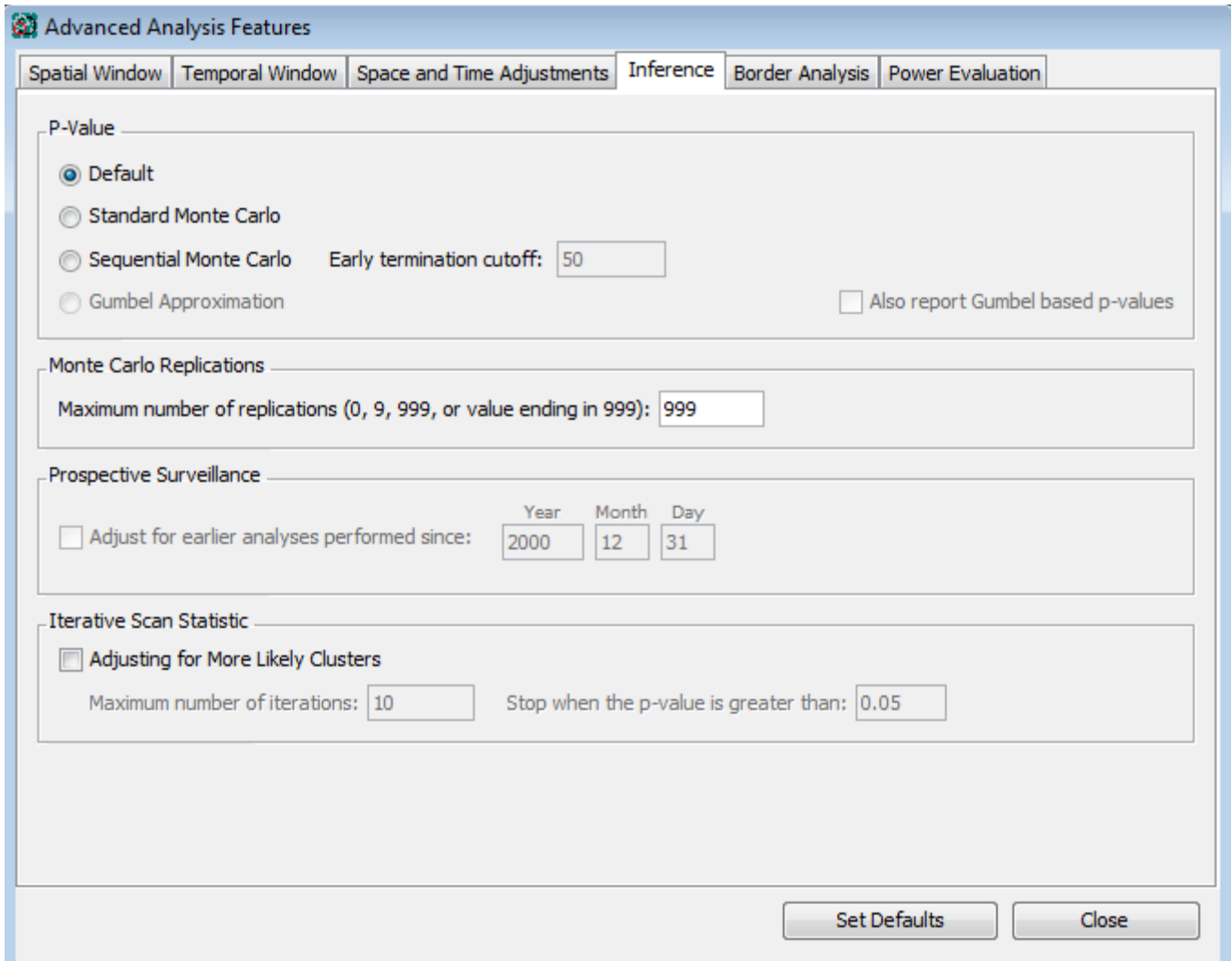


This opens a screen with six tabs. We are concerned with the first of these, 'Spatial Windows'. The default in SaTScan is to look for clusters covering up to half the population at risk, or to be more precise, half of the total expected counts. In New York State, that can be a very large area, often resulting in clusters containing all of New York City (about 40% of the state's population), the entire state north of New York City (another 40% of the population), or all of Long Island (20%). These can be interesting findings, but usually we are interested in identifying more focused areas. To avoid the detection of such large clusters, a smaller maximum cluster size can be set. We will choose a maximum of 10% of the population at risk.



We will look at one additional tab, '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).

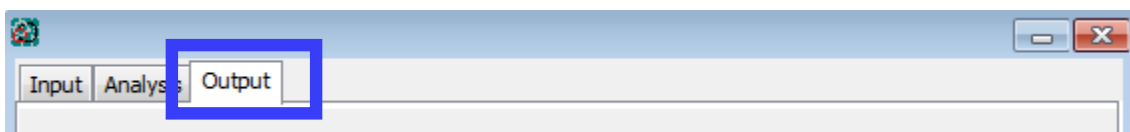
Sometimes, with very large data sets, SaTScan can take a long time to execute. In these situations, it can be useful to change the number of replications to 0 or 9 to make sure that the parameters are correct before committing to the entire analysis. It is seldom useful to choose a value above 999 as very little additional precision is gained by doing more simulations. For the data set used in this tutorial, the program should execute quickly with 999 permutations.



We will ignore the remaining advanced analysis tabs for this analysis, and simply accept the default settings. Click on 'Close' in the bottom right corner. The 'Set Defaults' button, here and on any other screen where it appears, will restore the default settings for all of the tabs.

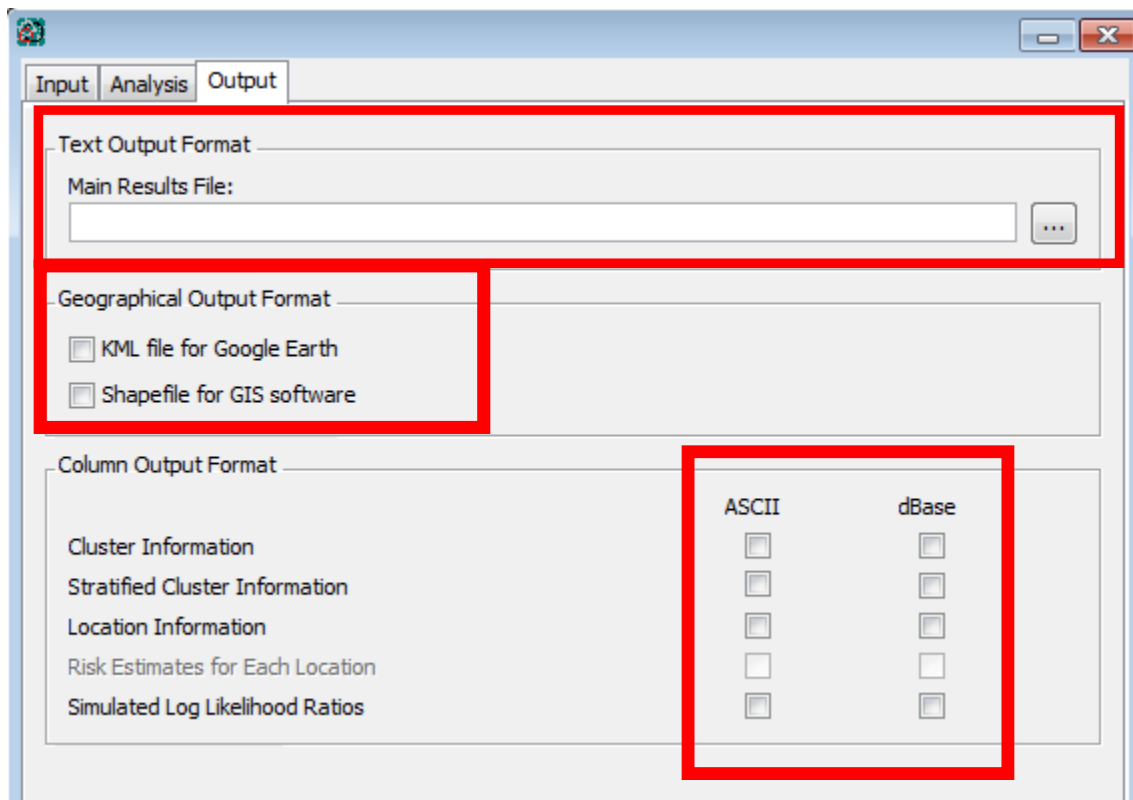
## 8. Output Tab


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.

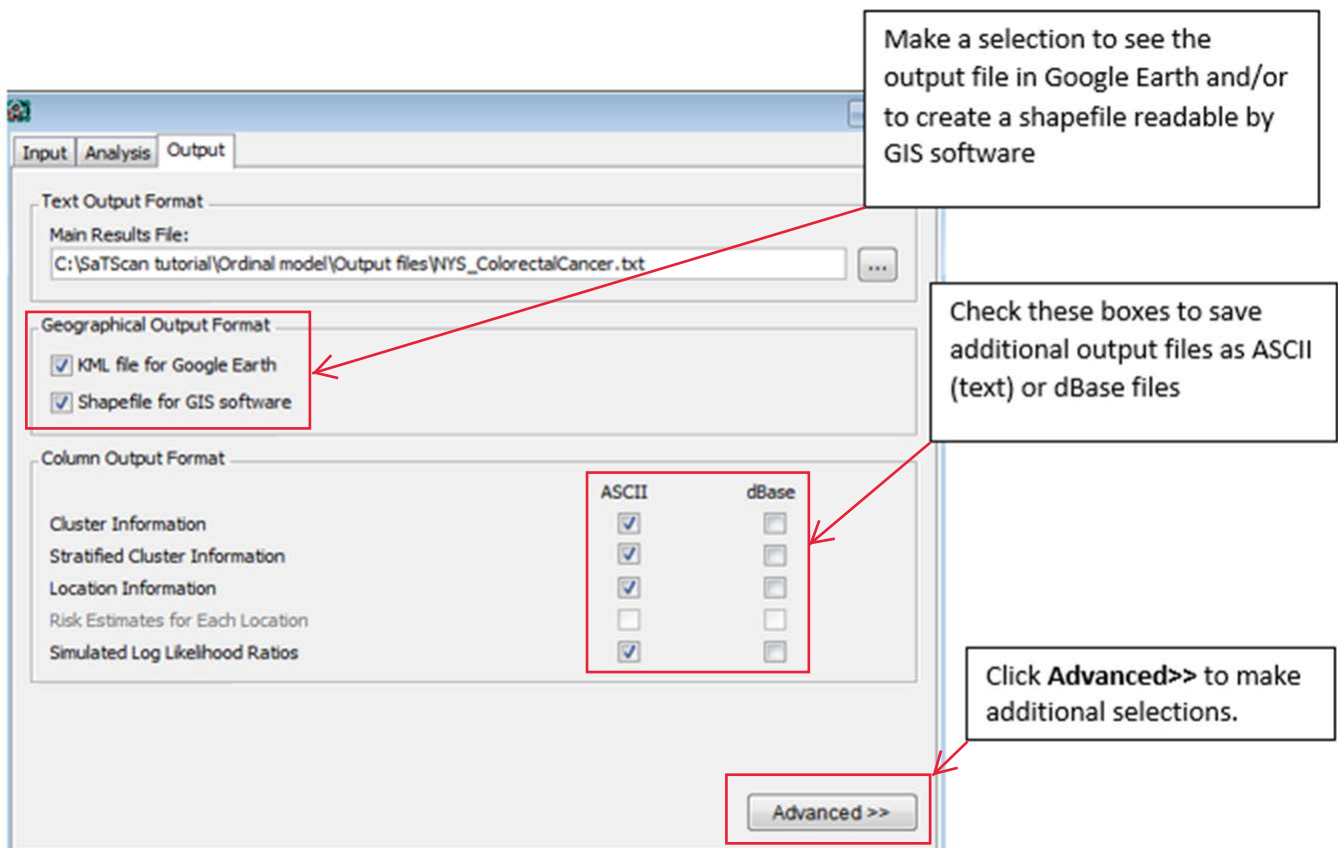
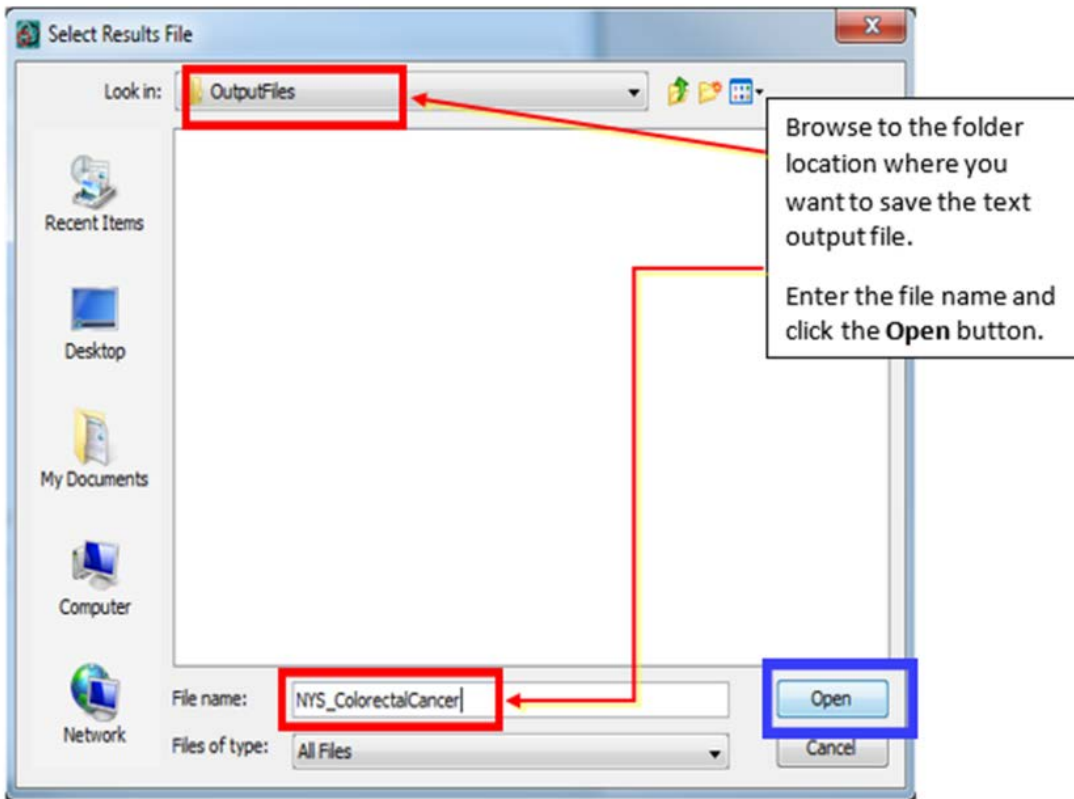




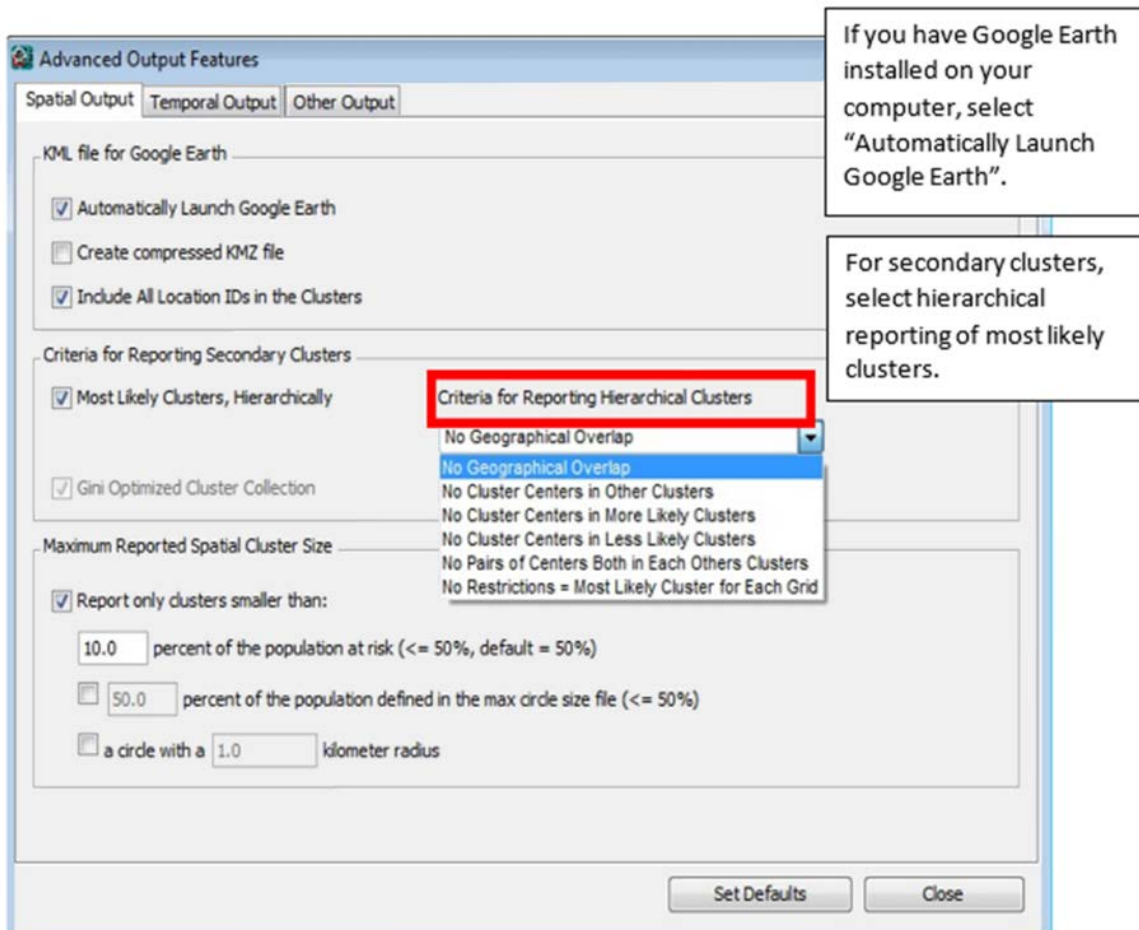
There are three sections on the main output tab corresponding to text, geographical, and column output formats. The Results File may be read directly in a text editor or word processing program, while the Column Output Formats are designed to function as input files to other computer programs for further display or examination of the results. The Geographical Output Formats are specifically designed for geographical display using geographical software such as QGIS or Google Earth.



Click on the  icon to modify/select the location of SaTScan output file. The file will be saved as a text file and will include 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.



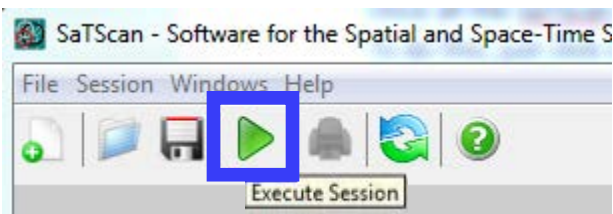
SaTScan will typically find multiple overlapping clusters, most of which are nearly identical to each other. These can be filtered in the 'Criteria for Reporting Secondary Clusters' section of the Advanced Output Features. For this tutorial, select 'No Geographical Overlap', which is the most restrictive 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.



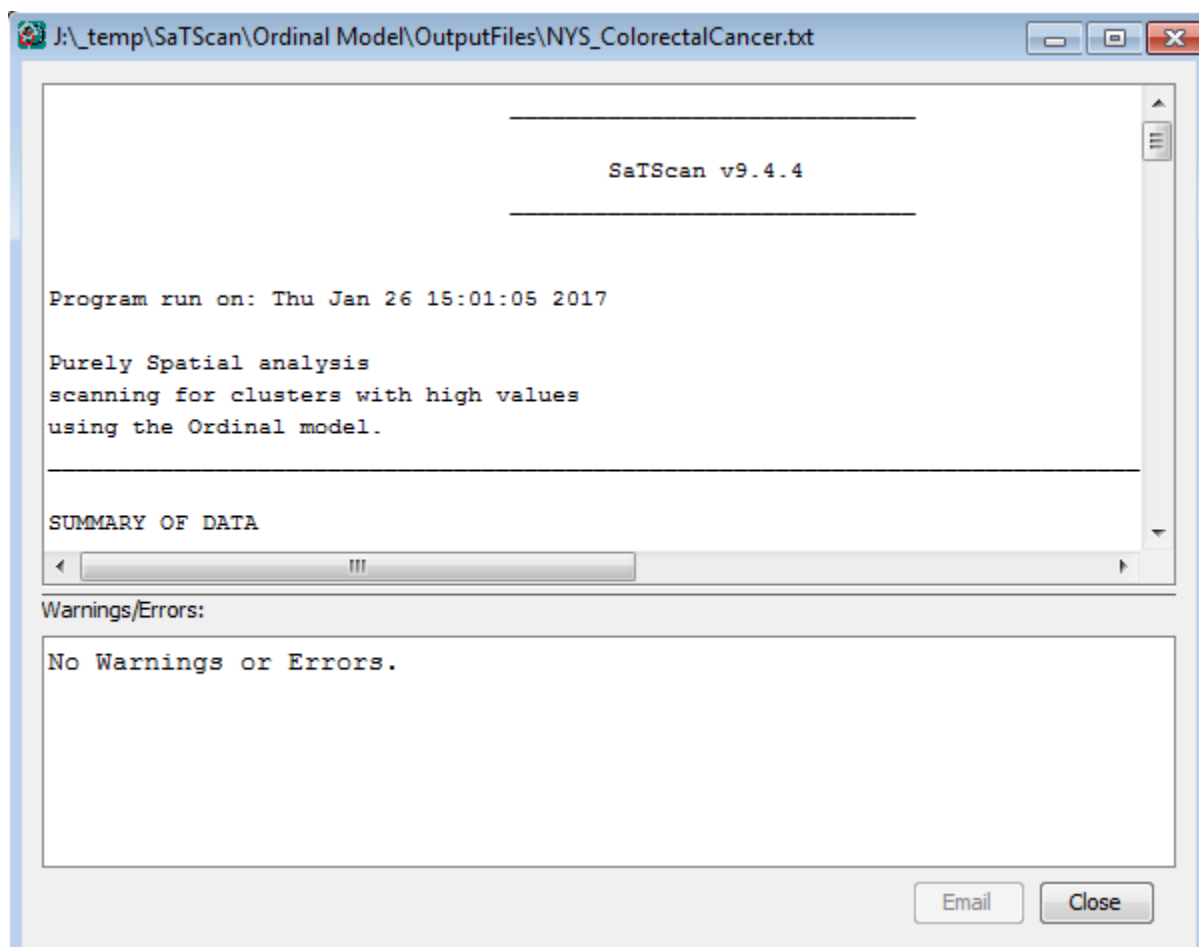
At last, it's time to run the analysis.

## 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. Once the analysis is finished, this window will show the contents of the results file, the same file specified in section 7. 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 indicates the type of analysis, the probability model and whether the analysis was used to scan for high, low or both high and low rates. If you ran the analysis as intended, it should say: “Purely Spatial analysis scanning for clusters with high values using the Ordinal model.” All other parameter settings are listed at the end of the Results file.

Next comes a summary of the data. You can use this to check that the data you have analyzed were the right data. If all went well, it should look like this:

#### SUMMARY OF DATA

```
Study period.....: 2010/1/1 to 2014/12/31
Number of locations.....: 952
Total number of cases.....: 43792
Category values.....: 1, 2, 3, 4
Total cases per category.....: 10910, 11920, 12147, 8815
Percent cases in area per category.: 24.9, 27.2, 27.7, 20.1
```

We see there were 952 locations and 43,792 total colorectal cancers diagnosed from 2010 to 2014, classified into four categories. All of the summary data agree with the numbers originally given in section 3.

The next part of the results file contains what we are most interested in: 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. In the New York State colorectal cancer data, there was just one statistically significant cluster. Here is its description:

CLUSTERS DETECTED

```
1.Location IDs included.: 884, 867, 864, 868, 887, 851, 900, 836, 888, 855, 847, 895, 784, 856,
                        839, 875, 916, 892, 835, 880, 891, 871, 908, 915, 820, 931, 827, 896,
                        828, 764, 899, 883, 859, 924, 912, 819, 919, 831, 932, 815, 907, 860,
                        843, 904, 811, 863, 879, 943, 928, 927, 796, 740, 772, 803, 944, 936,
                        903, 823, 947, 911, 951, 807, 948, 940, 923, 935, 952, 950, 939, 704,
                        787, 945, 799, 949, 934, 795, 783, 942, 918, 756, 941, 872, 804, 791,
                        946
Coordinates / radius.: (40.868900 N, 73.845200 W) / 9.54 km
Total cases.: 3910
Category.: [1, 2], [3], [4]
Number of cases.: 1924, 1052, 934
Expected cases.: 2038.39, 1084.55, 787.05
Observed / expected.: 0.94, 0.97, 1.19
Relative risk.: 0.94, 0.97, 1.21
Percent cases in area.: 49.2, 26.9, 23.9
Log likelihood ratio.: 18.416053
P-value.: 0.001
```

This cluster is centered at 40.8689 north latitude and 73.8452 west longitude and is 9.54 kilometers in diameter, which covers the Bronx in New York City and some surrounding areas. The cluster contains 3,910 cases (note that this is less than 10% of the total, as we specified), and the stages have been reclassified into 3 groups – stages I and II combined, with a relative risk of 0.94; stage III, with a relative risk of 0.97, and stage IV, with a relative risk of 1.21. As the first two groups are similar and close to 1, the real story here is with stage IV – in the Bronx and surrounding areas, there is 20% more diagnosis at late stage than in New York State generally. This is consistent with previous research that the Bronx is a problematic area for colon cancer screening. The p-value of 0.001 indicates that this is a significant cluster. It means that the calculated log likelihood ratio of 18.416 exceeded the value generated in every single one of the 999 simulations.

Additional non-significant clusters are also reported. For example, here is a small cluster of only 184 cases with an unusually low proportion of stage I cases:

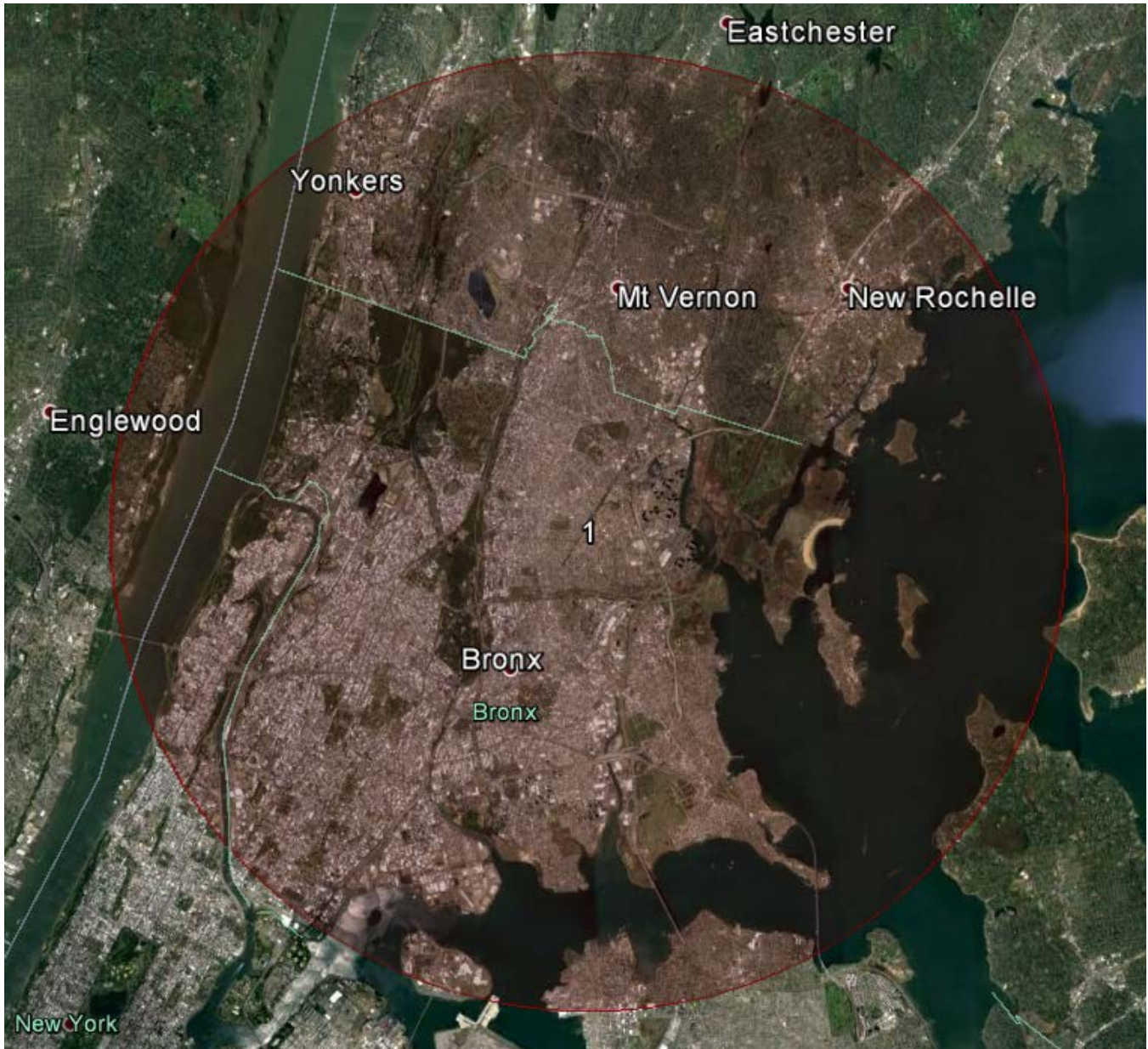
```
3.Location IDs included.: 116, 100, 128, 92
Coordinates / radius.: (40.907500 N, 73.062800 W) / 3.68 km
Total cases.: 184
Category.: [1], [2, 3, 4]
Number of cases.: 25, 159
Expected cases.: 45.84, 138.16
Observed / expected.: 0.55, 1.15
Relative risk.: 0.54, 1.15
Percent cases in area.: 13.6, 86.4
Log likelihood ratio.: 7.207973
P-value.: 0.853
```

However, since the reported p-value is 0.853, meaning that in 853 of the 1,000 simulations a result at least this unusual was obtained, this cluster is likely due to chance.

After the list of identified clusters, the results file contains details about all of the supplemental output files and parameters used in the analysis.

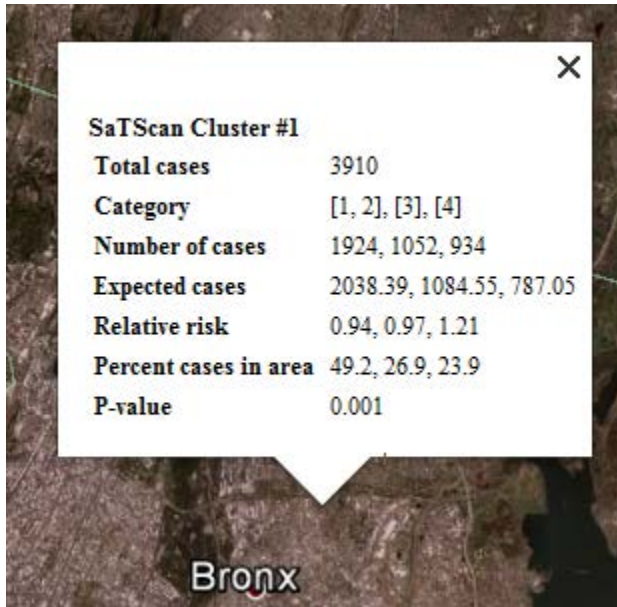
## 10.2 Google Earth View

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

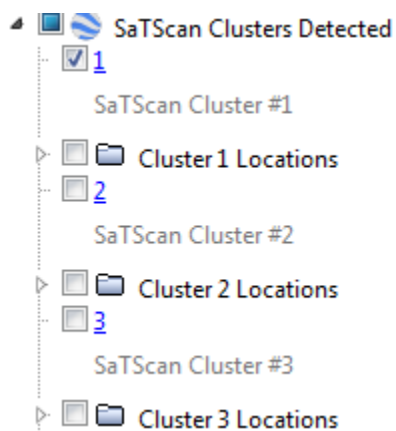


Your version of the map may have features like roads and places of interest visible – these can be turned on or off in the Layers window on the lower left of the screen.

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



The non-significant clusters are also included in the Google Earth output, but they are turned off by default. They can be made visible by checking them in the Places window at the left of the Google Earth screen. In the image below, only cluster #1 is visible:





### 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 to more easily import them 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 stage 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 results of the simulations. Regarding this last file, it shows the maximum log likelihood ratio generated in any of the simulations was 16.684 (obtained by opening the file in Excel and sorting from high to low). The value of 18.416 reported for the Bronx and surrounding areas is thus highly unlikely to have occurred by chance. (Since every simulation is different, you may see a different result than 16.684 but it should be close to this). Finally, the shape file (\*.shp) can be imported into GIS software for viewing on a map.

## 11. References and Further Reading

This is the fourth in a series of SaTScan tutorials. We also recommend reviewing the first three, which describe the purely spatial Poisson model using cancer incidence data, the Bernoulli model using birth defects data, and a second version of the Poisson tutorial that covers some of the advanced options within SaTScan. 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 ordinal model as it is implemented within SaTScan, we recommend the paper “A Spatial Scan Statistic for Ordinal Data” by Inkyung Jung, Martin Kulldorff and Ann Klassen, published in the journal *Statistics in Medicine* in 2007 (volume 26, pages 1594-1607). Please note that this paper is much more technical than the language used in this tutorial.