SaTScan Tutorial #2

The Bernoulli Spatial Scan Statistic for Birth Defect Data

Thomas Talbot and Sanjaya Kumar
New York State Department of Health

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

August 2015

v1.0

1. Introduction

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 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 the SaTScan™ software, when analyzing 0/1 case-control type binary data. As such, it complements SaTScan Tutorial #1, which covered Poisson-based disease incidence or mortality type data. This second tutorial can be utilized independently, without first having worked through SaTScan Tutorial #1.

This tutorial is intended for both self-learning and classroom use. 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 second SaTScan tutorial, we use the purely spatial scan statistic to analyze the geographical distribution of birth defects in New York State, USA, in order to determine if there are any geographical clusters of birth defects. That is, we will determine if there are any geographical areas with more birth defects cases than would be expected if the risk of birth defects was evenly distributed across the State.

The tutorial is written for SaTScan version 9.4 for Windows. The software tabs for subsequent versions may be slightly different than the screen shots shown in this tutorial, but they will be almost the same and there should not be a problem using the tutorial for subsequent versions. You can also use this tutorial if you use SaTScan for Linux or the Mac, except that some of the file handling steps will have to be adapted to those operating systems.

2. New York State Birth Defects Data

The data for this tutorial consists of birth defects in New York State, for the years 2005 to 2009. The data comes from the New York State Congenital Malformation Registry, which is a part of the New York State Department of Health. This is one of the largest statewide, population-based birth defects registries in the nation.

Birth Defects: Birth defects, also called congenital malformation, are serious conditions or changes to the structure of one or more body parts and it can affect almost any part of the body such as the face, brain, heart or limbs. The effect of birth defects on the expected life span depends on the severity as well as body part affected. According to the Centers for Disease Control and Prevention, birth defects affect 1 in every 33 babies born in the United States each year. Depending on the body part affected, the diagnosis of a birth defect can take time. Some birth defects are obvious at birth, such as cleft lip, whereas birth defects in internal organs can take longer and may need specific medical tests to detect. The Congenital Malformations Registry receives case reports on children diagnosed up to two years of age who were born or reside in New York State, and who were diagnosed with a congenital malformation, chromosomal anomaly or persistent metabolic defect.

Geographical Resolution: The data are provided at the level of the zip code. In order to protect patient confidentiality, a Zip Code needs to have a minimum of 100 births. zip codes not meeting this threshold were merged with neighboring zip codes. This resulted in a reduction in the number of geographical location units.

Birth Defect Data File: The birth defects data file for this tutorial can be downloaded from the SaTScan web site (www.satscan.org/datasets.html). It is a dBase file, with the name Birth_defcts.dbf. In addition to SaTScan, it can be opened in Excel and a large number of other programs. The file contains the following columns: zip code, year, number of births in that zip code and year, the number of births with a birth defect, the number of births without a birth defect, the latitude of the zip code and the longitude of the zip code. The data is also available directly from the New York State Department of Health, at New York State Health Data NY webpage in a different format, (health.data.ny.gov).

Data Dictionary: For the purpose of this tutorial we will need to **create the following data directory structure**. You can set it up differently in any location, but it is important to have at the three folders bolded, this will be important later on in the tutorial:

Main Directory   C:\Temp\SaTScan**Bernoulli Model**
SaTScan Input Files   C:\Temp\SaTScan\Bernoulli Model**InputFiles**
SaTScan Output Files   C:\Temp\SaTScan\Bernoulli Model **OutputFiles**
*File required for tutorial*: Birth_defects.dbf

3. SaTScan Software Download and Installation

 To download the free SaTScan software, please go to www.satscan.org. Select "download", and follow the instructions. In order 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 detected clusters on Google Earth. For this feature to work, you need to install Google Earth on your computer. To download, go to http://www.google.com/earth/index.html, and follow the instructions. This is an optional step, and you can run SaTScan perfectly well without Google Earth.

4. Launching the SaTScan Software

Launch the SaTScan software by using one of the following three methods. Your set-up may vary slightly on your computer.
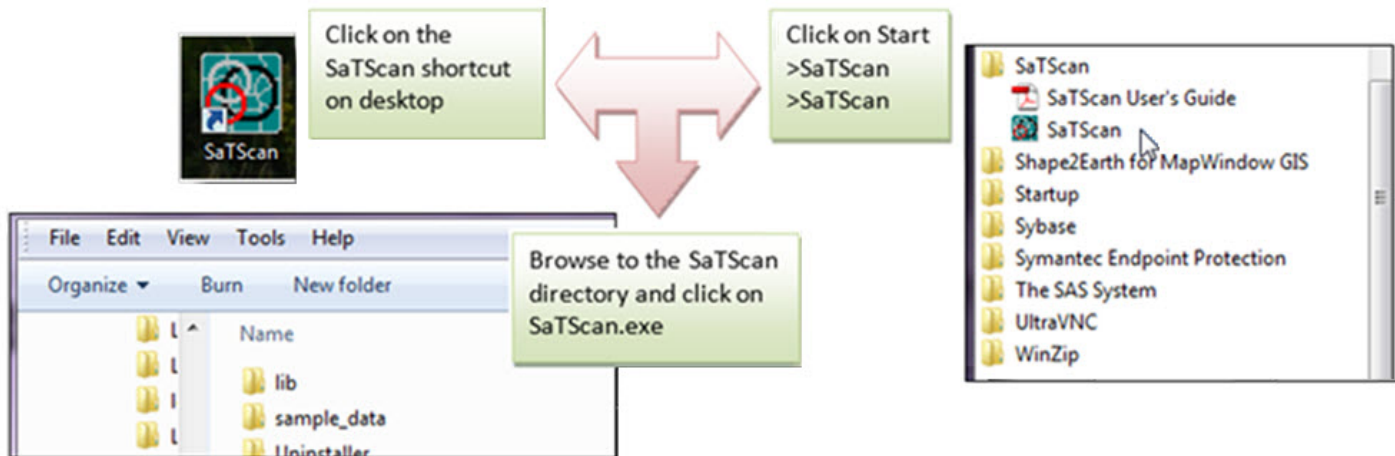
As the next step, select "Create New Session" from the Start Window.
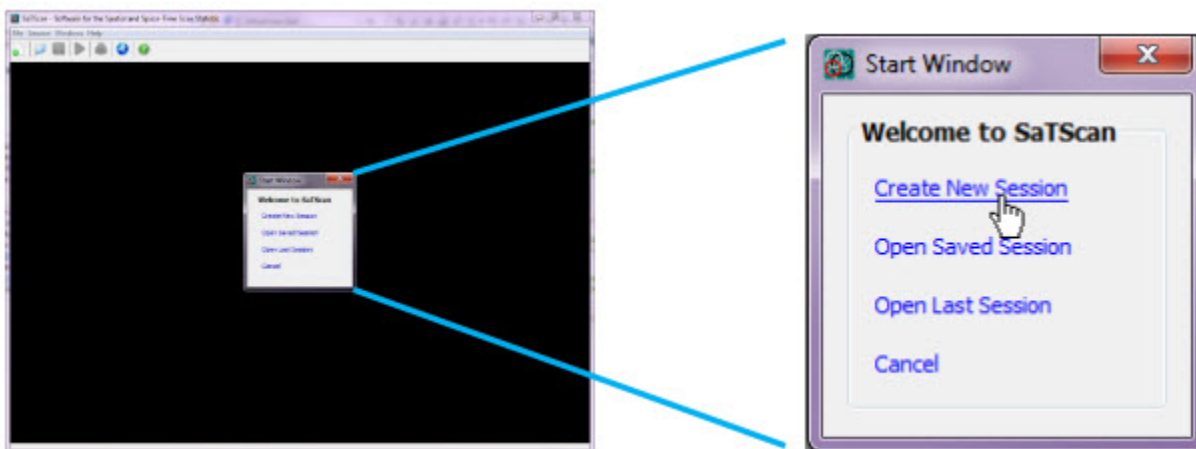
You should now see the input data tab, and you are ready to specify the analysis you want to run. The SaTScan software has three main tabs for specifying the input data, analysis parameters and output formats respectively. We will go over each in turn. Each of the three main tabs also have a set of advanced tabs to define less used software options.

5. Input Data Tab

The first of the three main tabs is used to specify the input data. For 0/1 binary data and the Bernoulli model, which we will be using, three input files are required: a case for the birth defect cases, a control file for all birth without a birth defect and a coordinate's file for the geographical latitude/longitude information. These three files will all be created from the Birth_defect.dbf file that you already should have downloaded from the SaTScan web site. The three files are linked through a location ID, which represents the name or code for some geographical entity such as a state, province, county, zip code, postal code area, census tract or dwelling. In the New York State data, the locations used are zip code, with each one represented by a set of five integers.

5.1 The Case File (*.cas)

If the Case File is 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, we do not have the birth defect cases in such a format, so we will instead use the SaTScan Import Wizard, in order to create the case file from the Birth_defects.dbf file.
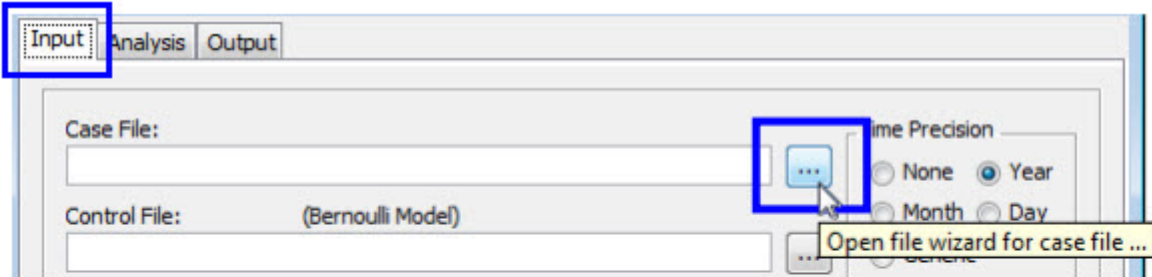
Click on the ⸚ **button** to import the case data using the SaTScan file wizard.

Select Birth_defects.dbf 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 *.dbf file.
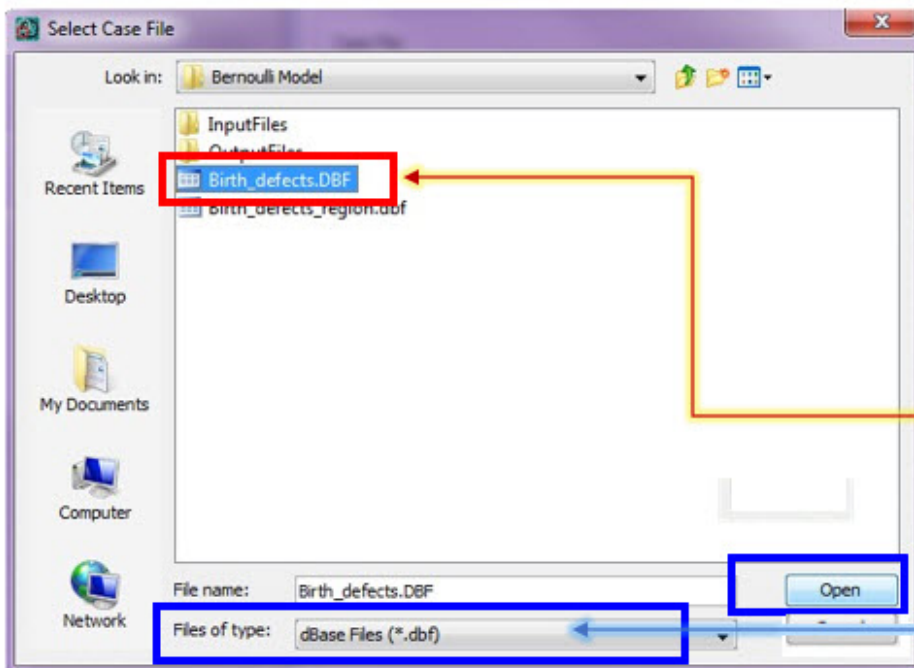
On the 'Select Case File' dialog:

Change 'Files of Type' to dBase Files (*.dbf)

Select Birth defect.dbf as the input file.

Click **Open**
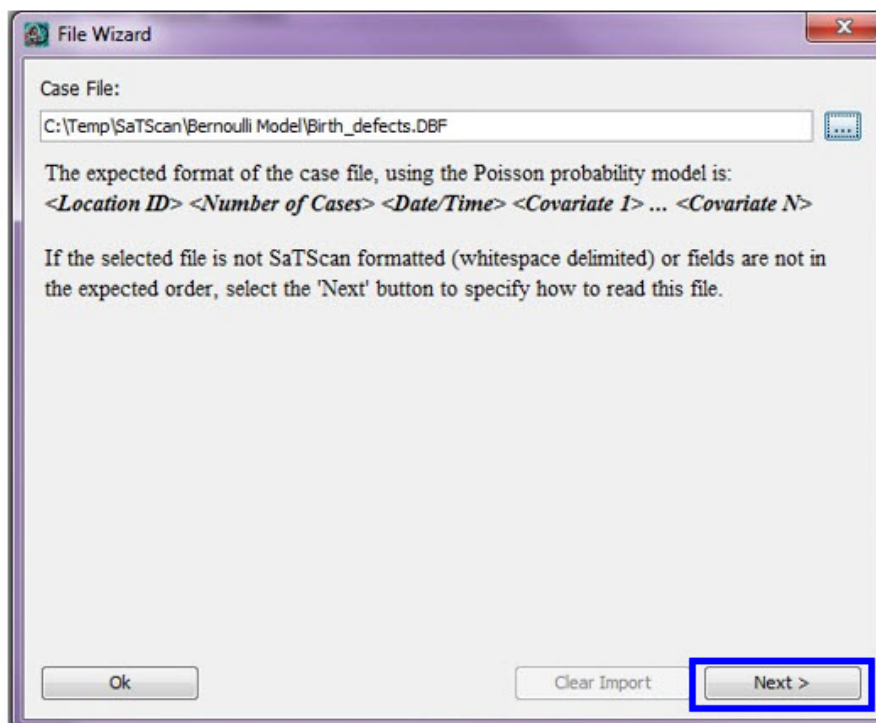
On the 'File Wizard' dialog, check the input case file and click the Next > button.

Once the import file has been selected, and you have clicked the 'Next >' button, you must assign columns to each of the required SaTScan variables. Since we are going to use the Bernoulli probability model, you need to tell the wizard displays the SaTScan variables for that model.

Change
Set the *Display SaTScan Variables For* option to **Bernoulli** model.

Assign
Column names for *Location ID, Number of Cases,* and *Date/Time.*
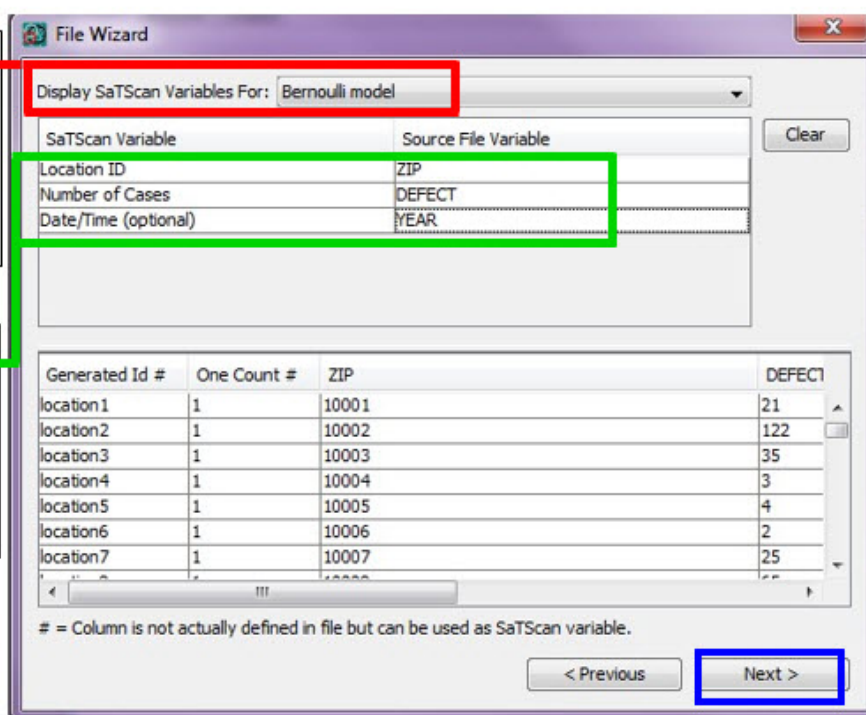
For this tutorial, we suggest that you change the file name to NYS_BirthDefects.cas. This file is in the SaTScan case file format, with the extention *.cas. It is automatically assigned as the case file for the current analysis, but it can also be used subsequently without having to go through the import wizard process again.

After assigning the variables, click on next. You are now asked to specify the file name and the directory in which you want to save the SaTScsan case file. You do not have to, but if you want to, you can change the name of the case input file to something that is easier to remember.

1. Click on the **Change** button and browse to the **Input Files** directory of the training folder and

2. Append the file name **NYS_BirthDefects.cas**. **NOTE:** When you change the file location, the software assigns the default name **Cases.cas**, so you should rename it
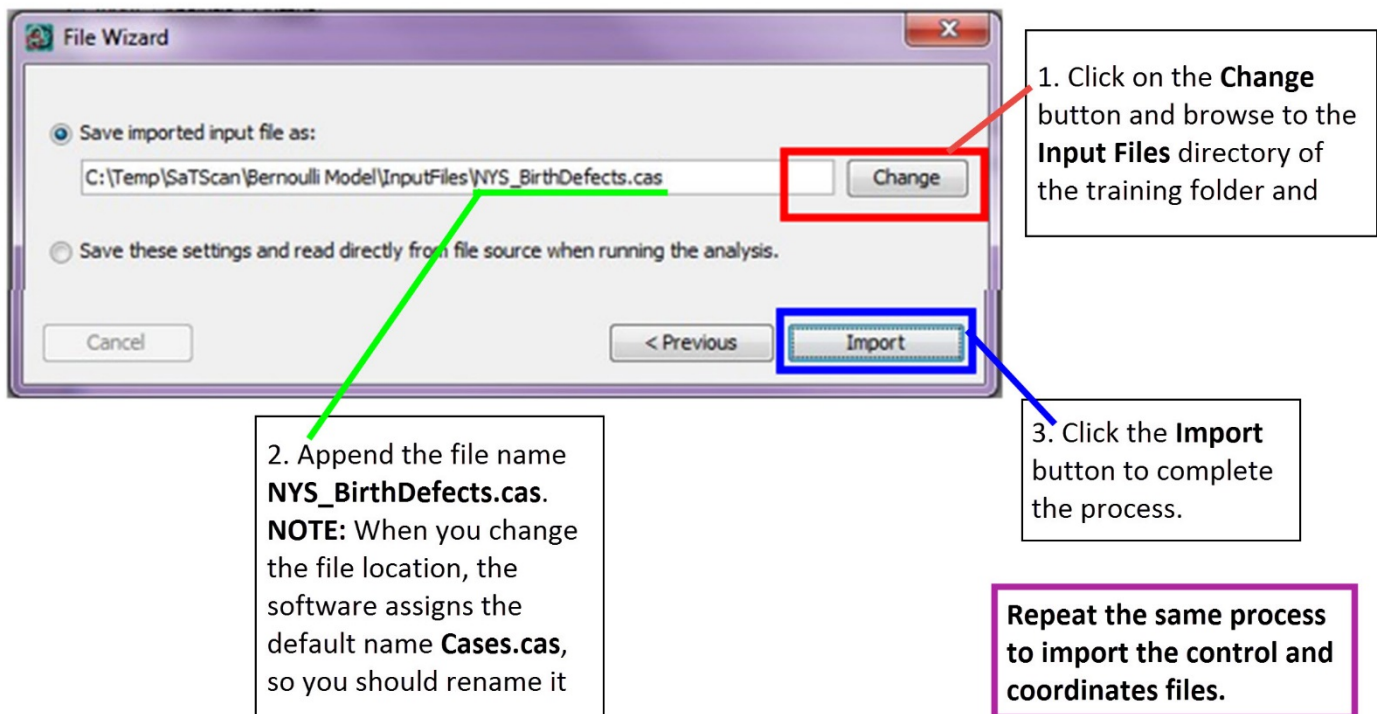
3. Click the **Import** button to complete the process.

Repeat the same process to import the control and coordinates files.

5.2 The Control File (*.ctl):

The Bernoulli model is designed for data containing individuals who do or do not have a disease and for other 0/1 type variables. For the Bernoulli model, the controls could be a random set of controls from the population or the total population except for the cases. In the birth defects data, the controls are calculated by subtracting the number of births with birth defects from the total number of live births. These are listed in the Birth_Defects.dbf file under the NODEFECT column.

Control File: (Bernoulli Model)
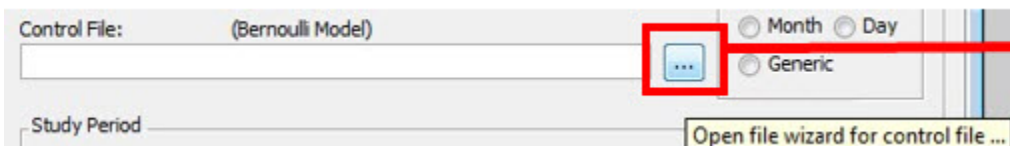
To begin open the **File Wizard**.

Open file wizard for control file ...

Study Period

Select Control File

Look in: Bernoulli Model

InputFiles
OutputFiles
Birth_defects.DBF
Birth_defects.ID
Birth_defects.MAP
Birth_defects.TAB
Birth_defects_region.dbf
Birth_defects_region.prj
Birth_defects_region.shp
Birth_defects_region.shx

Next, select the file, **Birth_defects.dbf**.

File name: Birth_defects.DBF

Files of type: All Files

Open

Cancel

Click the **Open** button.

File Wizard

Control File:

C:\Temp\SaTScan\Bernoulli Model\Birth_defects.DBF

The expected format of the control file is:
<Location ID> <Controls> <Date/Time>

If the selected file is not SaTScan formatted (whitespace delimited) or fields are not in the expected order, select the 'Next' button to specify how to read this file.

Ok   Clear Import   Next >

Click the **Next** button.

Click on **Change** and change the file name from Controls.ctl to NYS_BirthDefects.ctl

5.3 The Geographical Coordinates File (*.geo)

We cannot do a spatial analysis without information about the spatial location of the birth defect cases and all other births. Using the same location IDs as in the case and control files, we have to specify the geographical coordinates for each location ID. That is done using the coordinates file. As above, we will use the SaTScan Import Wizard.

Check the selection in the 'Display SaTScan Variables For' dropdown list. There are two options available, and you should make selection based on the coordinate system of your input dataset. For this tutorial, we will select "**Latitude/Longitude Coordinates**" as shown in above screen image.

Make selections for SaTScan variables using the dropdown list under the Source File Variable column. You will need to make a selection for each SaTScan Variable as shown.

Click on **Change** and change the file directory to Input Files directory file name from Coordinates.pop to NYS_BirthDefects.geo.

Add the file name for the geographical file as NYS_BirthDefects.geo
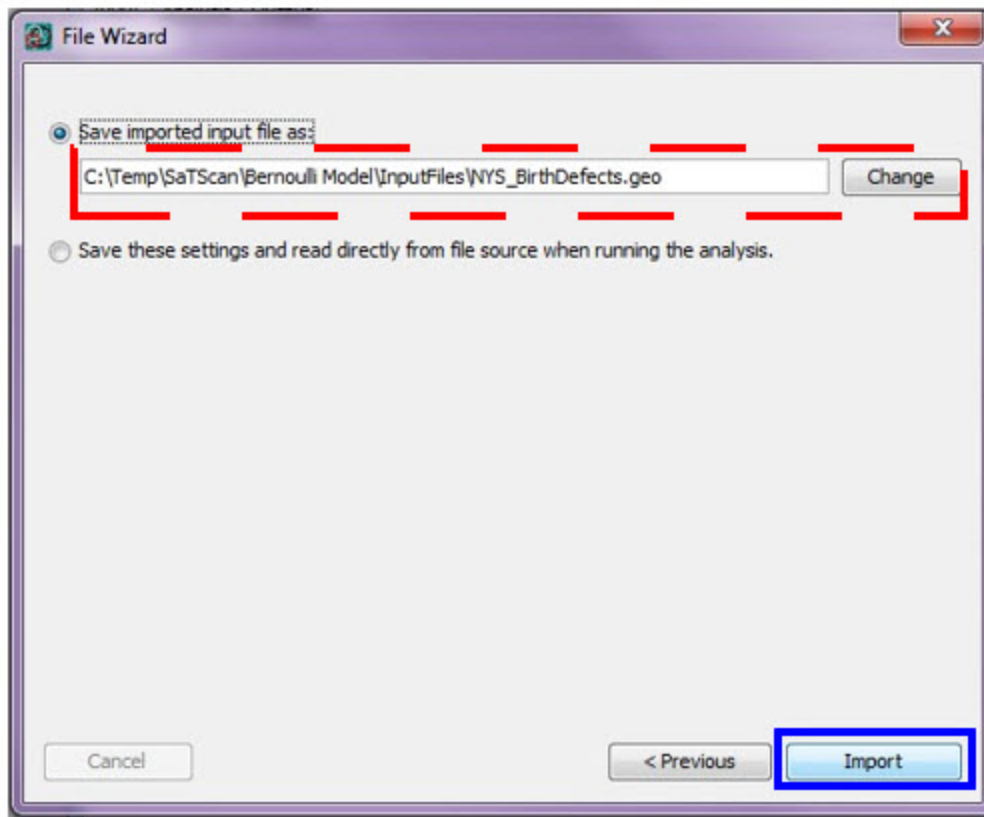
Click Import to create the NYS_BirthDefects.geo file.

5.4 Information about the Input Data

After specifying the names of the SaTScan input files, it is necessary to provide SaTScan with some information about the content of the new files. This is also done on the main input tab.

Time Precision (input tab)
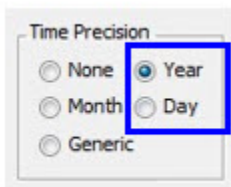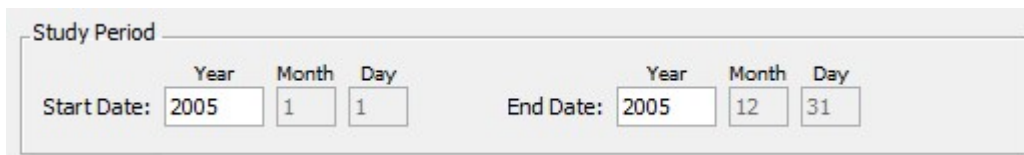
The first piece of information to specify is the precision of the times specified in the case file. If there are no times in the case file, "none" is selected. Using integer values, generic is used for times not specified using calendar dates. In the New York birth defects data that we imported, time is given as the year of the incident diagnosis, and hence we select "Year".

For a purely spatial analysis, it is not necessary to have a time variable, but it is necessary to specify whether the time variable is present in the imput case file or not, in order to read the file correctly.

Study Period (Input tab)

The next step is to define the study period. This must be done even for a purely spatial analysis, when there are no times in the case or control files, in order to properly calculate annual disease rates. In the Birth_Defect.dbf file, this is the time period for which we have collected births and birth defects, from 2005 to 2009. Hence, we select January 1, 2005 as the start date and December 31, 2005 as the end date.

Coordinates (Input tab)

In SaTScan, the geographical locations can either be specified as latitude and longitude, or, as Cartesian coordinates, which is the regular x,y-coordinate system taught in high school. For the New York birth defect data, we have latitude and longitude coordinates. It is important to make the correct selection, since otherwise, the intended circles will not be circular. Note that, when latitude/longitude is used, SaTScan draws perfect circles on the surface of the earth, and no map projection is used.

5.5 Final Check of Input Tab

You should now be done with the input tab, but before moving on to the next tab, it is a good idea to double check all the entries to make sure they are right. To do so, you can compare your input tab with the one below.

6. SaTScan Analysis Parameter Settings

The first choice on the analysis tab is the type of analysis. In this tutorial, we plan to do a *purely spatial* analysis, ignoring the time dimension.

Next, it is necessary to choose the *probability model*. We have binary 0/1 count data, where a birth either was or was not associated with a birth defect. For such data, it is suitable to use the Bernoulli probability model.
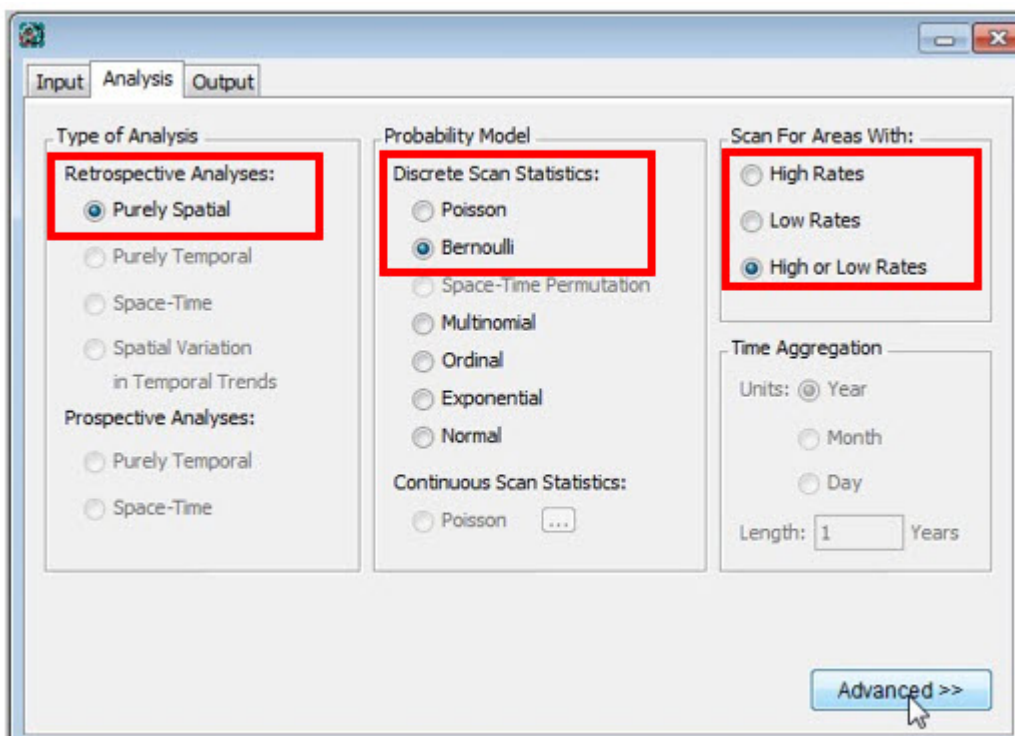
Spatial scan statistics are typically used to detect clusters of cases; that is, areas with a larger number of cases than would be expected by chance. This indicates areas where there may be a higher risk for the disease. Sometimes it is also of interest to look for areas with fewer cases than expected, where the risk of the disease is lower. In our case, let's say that we are interested in areas of either high or low risk for birth defects. We then select to scan for areas with high or low rates.

The last option on the Analysis Tab is for *Time Aggregation*, but that is only relevant for purely temporal and space-time analyses. Since we are doing a purely spatial analaysis, this option is greyed out, and we simply ignore it.

If we wanted to, we could now be done with the Analysis Tab and move on to the Output Tab, but we will instead go to the advanced tabs. This is reached by clicking on "Advanced" in the bottom right corner of the Analysis Tab. See the blue box below.

This image shows how to open the Advanced Analysis section from the the Analysis tab.
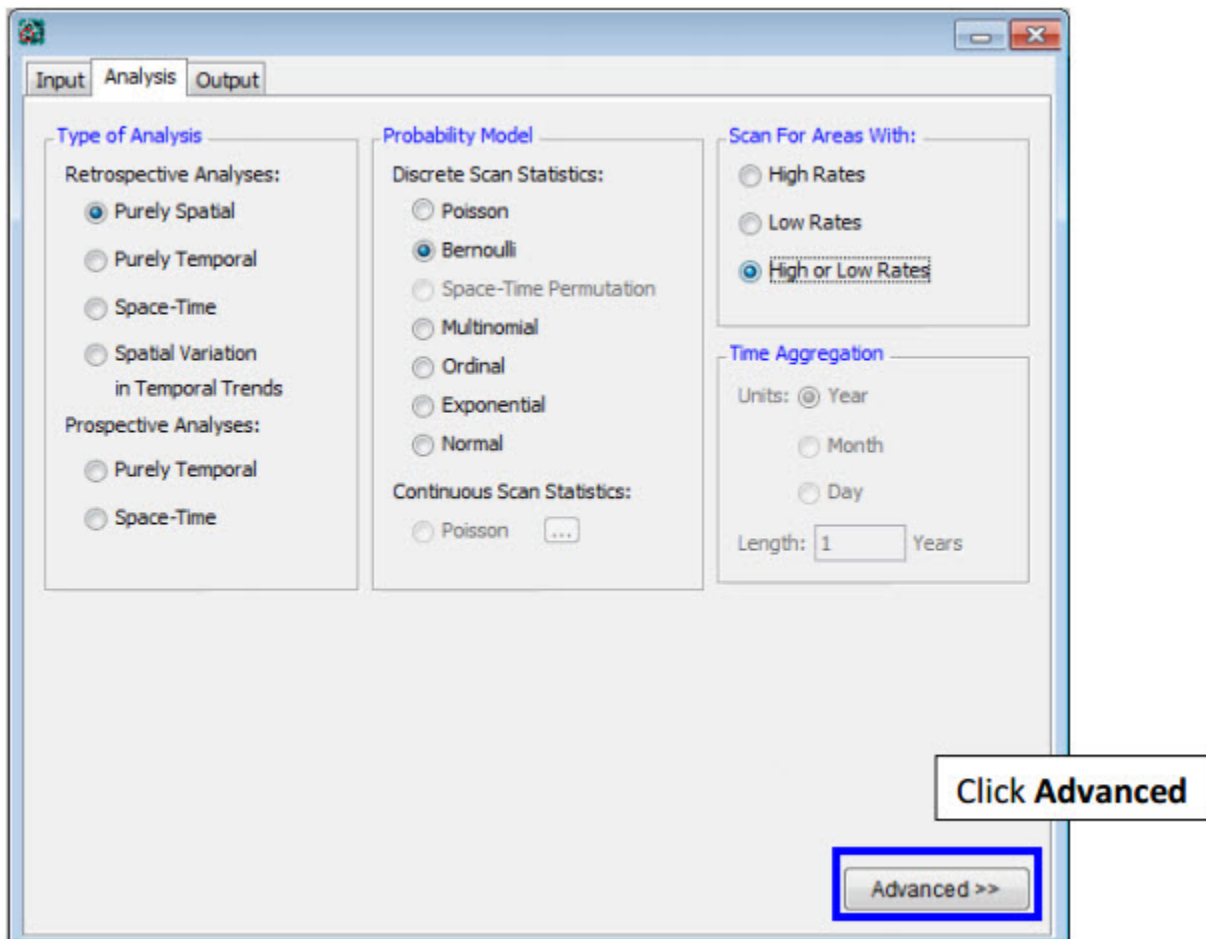
The first of the six advanced analysis tabs is called 'Spatial Windows', which is the one we want. The default in SaTScan is to look for clusters covering up to half the population at risk, or to be more precise, half of the total expected counts. In New York State, that can be a very large area, containing almost the whole state except New York City. To avoid the detection of such large clusters, one can set a smaller maxima on the cluster size. We will pick a maximum of 25% of the population at risk, and this is specified on the top of the 'Spatial Windows' tab.
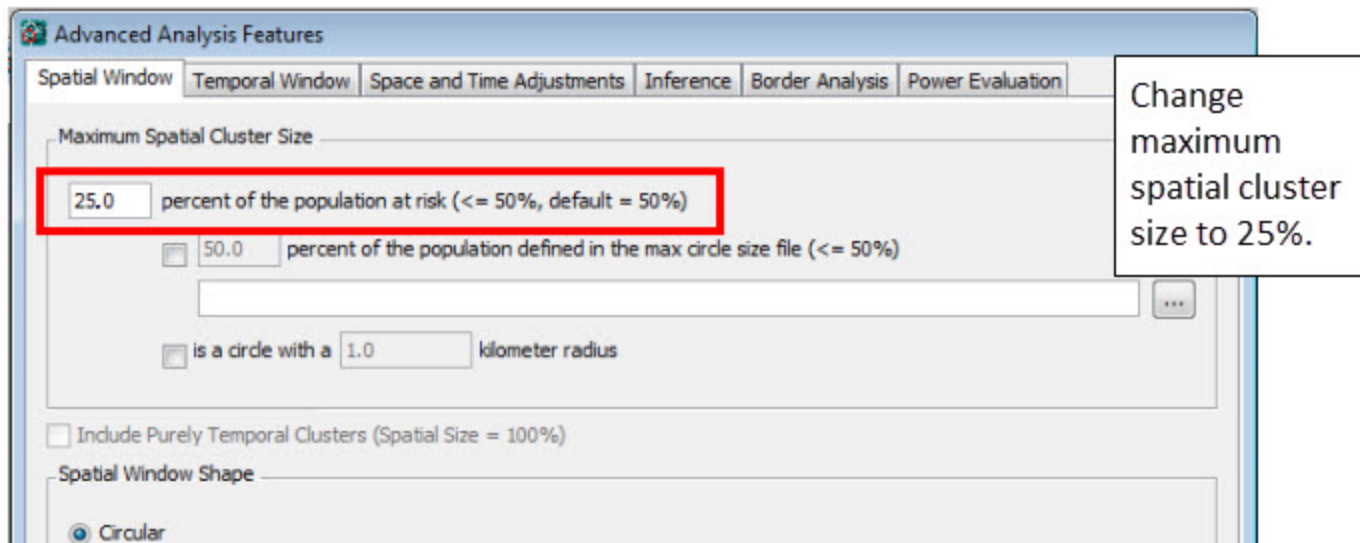
Note: It is inappropriate to run multiple SaTScan analyses with different maximum values. With a 25% maximum, both small and large clusters are evaluated, and SaTScan adjusts for the multiple testing inherent in the 0-25% range used. If multiple maxima are used, such as 5%, 10% and 25%, we are evaluating cluster over the whole 0-25% range, and we need to adjust for that multiple testing appropriately, and it is only the analysis with the 25% maximum that does that.

While there are many other advanced options, we will ignore those for our analysis, and simply use the default settings. Once all the advanced parameter setting have been specified, click on 'Close' in the bottom right corner. If you did some other changes, that you now regret, you can click on 'Set Default'. SaTScan will then erase your chosen analysis setting, on all the five advanced analysis tabs, and return them to the default values.

7. Specifying SaTScan Output Options

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

Output Tab:
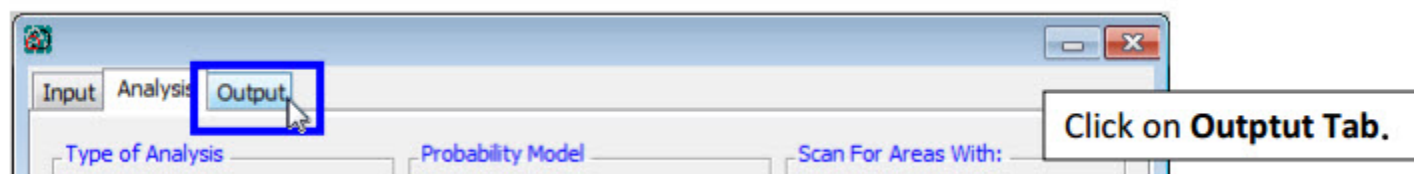
There are three sections on the main output tab for text, geographical, and column output format. The Results File is a compulsory output file in ASCII text format designed for the human eye, while the Column Output Formats are designed to function as input file to other computer programs for further display or examination of the results. The Geographical Output Formats are specifically designed for geographical display using geographical software such as GIS systems or Google Earth.
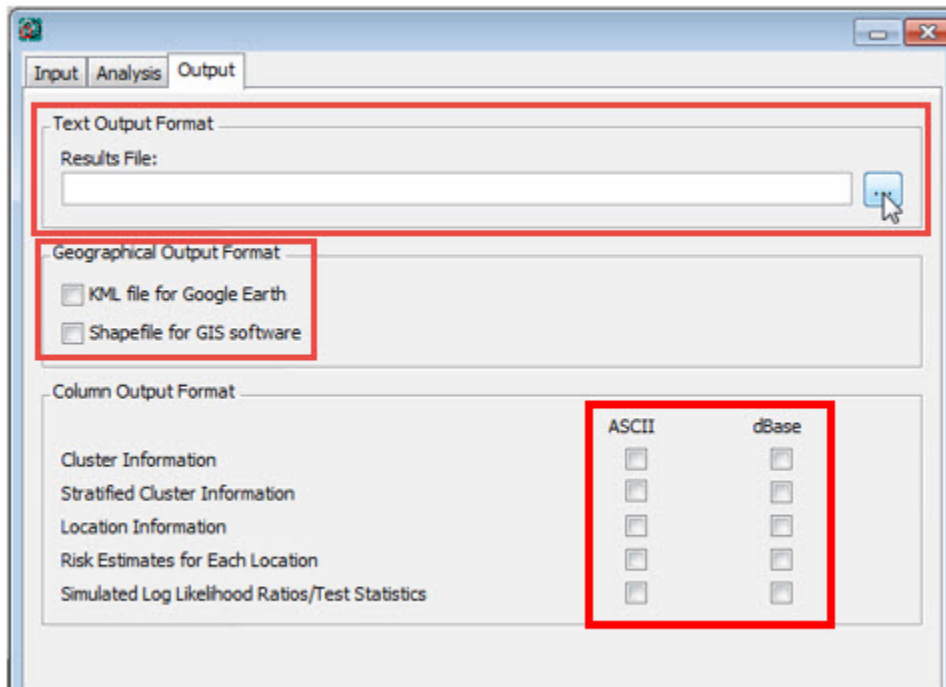
Click on the [...] icon to modify/select the location of SaTScan output file. The file will be saved as a text file and will have several important sections including a summary of the data, locaction IDs of each location included in each cluster, coordinates and radius of each cluster, population, number of cases, number of expected cases, relative risk and p-value for each cluster detected.

Select Results File

Look in: OutputFiles

Recent Items
Desktop
My Documents
Computer
Network

File name: NYS_BirthDefects

Files of type: All Files
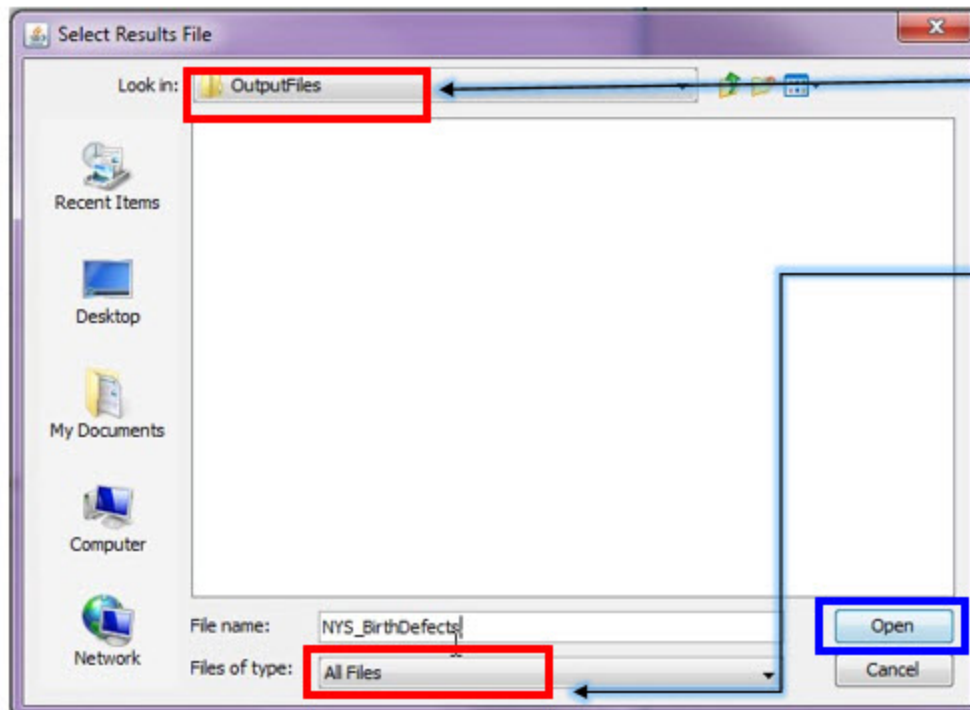
Open
Cancel

Browse to the folder location where you want to save the text output file.
Enter the file name and click the **Open** button.

Input | Analysis | Output

Text Output Format

Results File:
C:\Temp\SaTScan\Bernoulli Model\OutputFiles\NYS_BirthDefects.txt

Geographical Output Format
☑ KML file for Google Earth
☑ Shapefile for GIS software

Column Output Format

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

Advanced >>
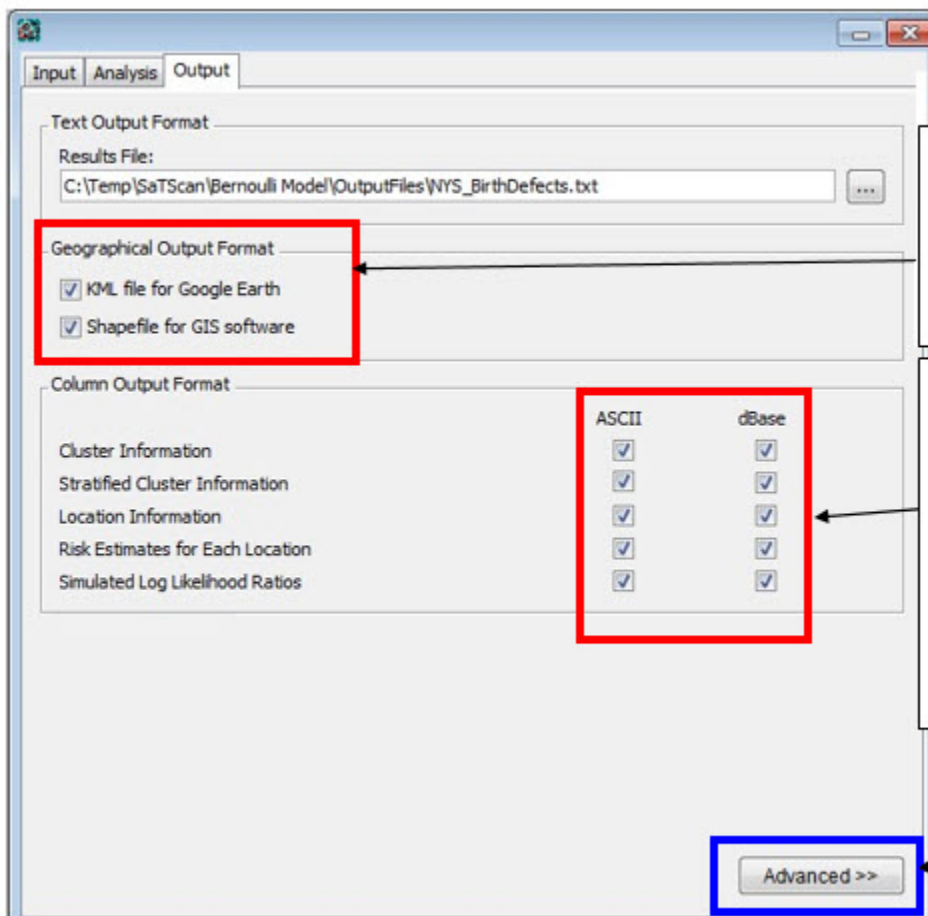
Make a selection to see the output file (clusters) in Google Earth and/or that can be imported into GIS

Make selections to save additional output files separately in one or more formats. You can select one or both formats. The ASCII files can be opened in Excel or other spreadsheet software.

Click **Advanced>>** to make additional selections.

Advanced Output Features

Spatial Output | Temporal Output | **Other Output**

KML file for Google Earth

☑ Automatically Launch Google Earth
☐ Create compressed KMZ file
☑ Include All Location IDs in the Clusters

> If you have Google Earth installed on your computer, select "Automatically Launch Google Earth".

Criteria for Reporting Secondary Clusters

☑ Most Likely Clusters, Hierarchically

Criteria for Reporting Hierarchical Clusters

No Geographical Overlap

No Geographical Overlap
No Cluster Centers in Other Clusters
No Cluster Centers in More Likely Clusters
No Cluster Centers in Less Likely Clusters
No Pairs of Centers Both in Each Others Clusters
No Restrictions = Most Likely Cluster for Each Grid Point

☐ Gini Optimized Cluster Collection

Maximum Reported Spatial Cluster Size

☑ Report only clusters smaller than:

25.0 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

> For secondary clusters, select hierarchical reporting of most likely clusters, deselecting the Gini optimized cluster collection.

[Set Defaults]   [Close]

SaTScan will typically find multiple overlapping clusters, some of which are almost identical to each other. In SaTScan there are options to report a different number of these overlapping clusters based on different criteria. For this tutorial, select 'No Geographical Overlap', which is the most restrictive choice. With this option, a secondary cluster will only be reported if it does not overlap with a more likely and previously reported cluster.

For the purpose of this tutorial, make selections as shown in screen below.

After making the selections, click on the **Other Output** tab at the top of **Advanced Output Features**.

If you have Google Earth installed on your computer, select "Automatically Launch Google Earth".

For secondary clusters, select hierarchical reporting of most likely clusters, deselecting the Gini optimized cluster collection.
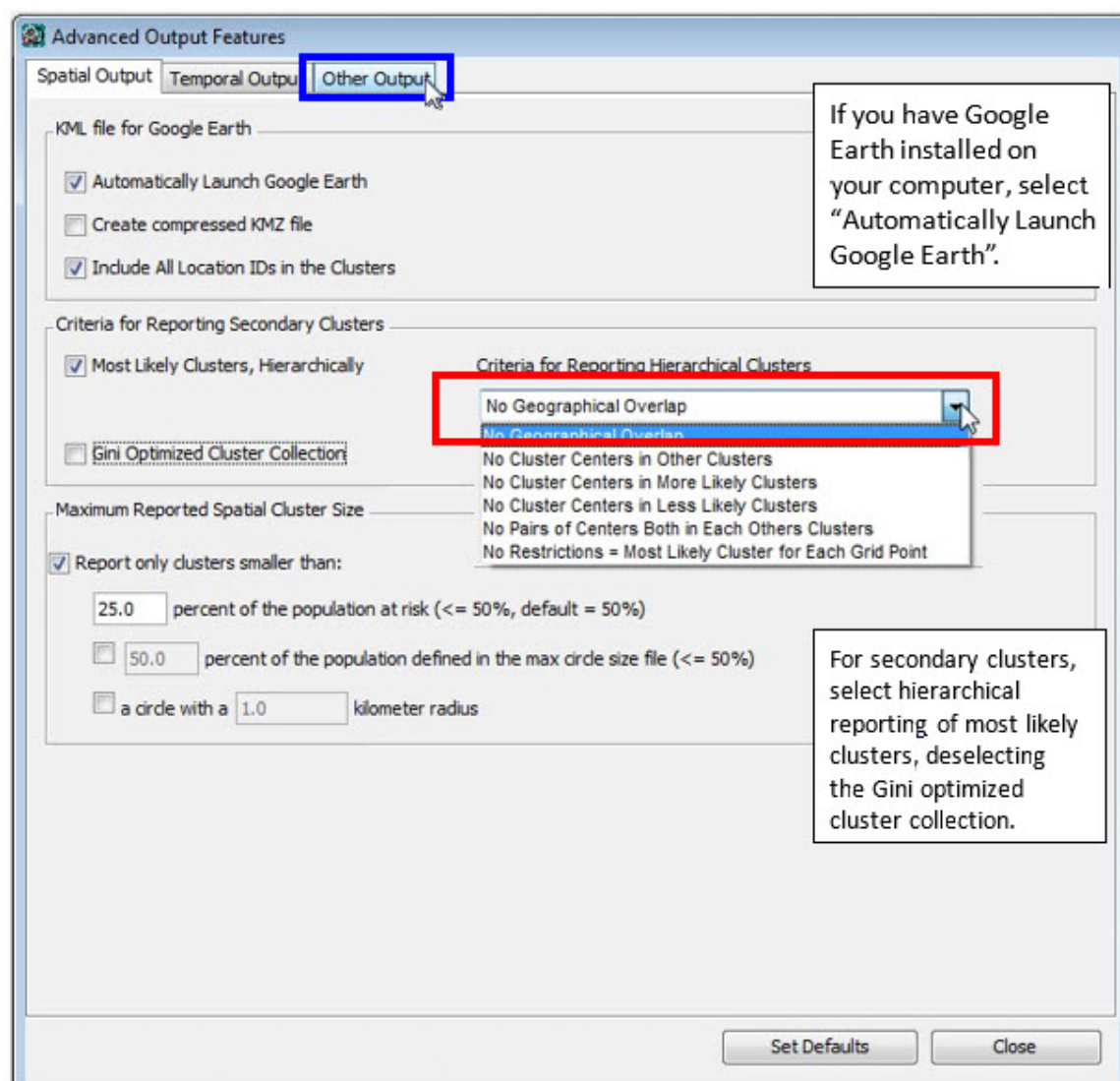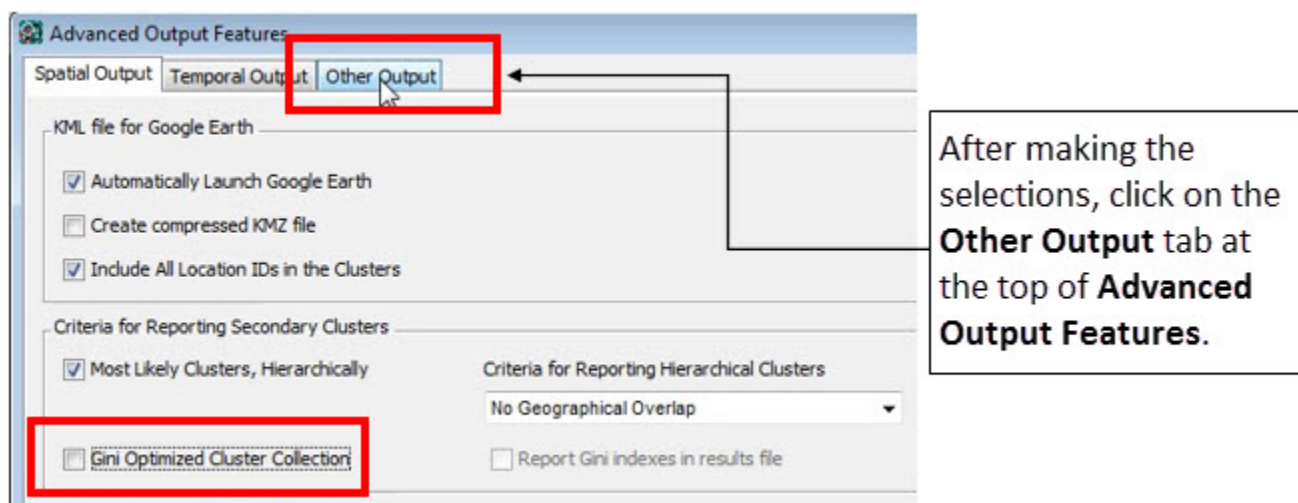
After you have made all the selections on the output tabs, you are ready to run a SaTScan analysis.

8. Running SaTScan

Once all the SaTScan parameters have been selected on the various tabs, it is time to run the program. To do this, just click on the button with the green triangle.
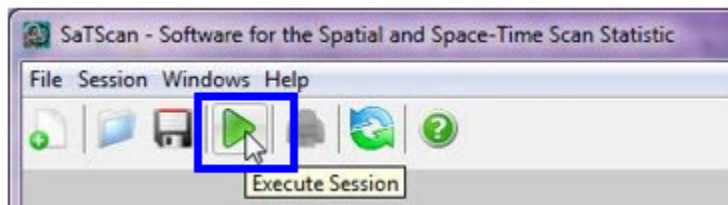
Once the SaTScan analysis is running, you will see a window that shows the progress being made. There is nothing in this window that you need to write down or memorize. After the analysis is done, this window will show the text based results file. You will be able to scroll through this window to see all the results, as well as the parameter settings you used. These are the same results that are also saved in the file that you specified in section 7.
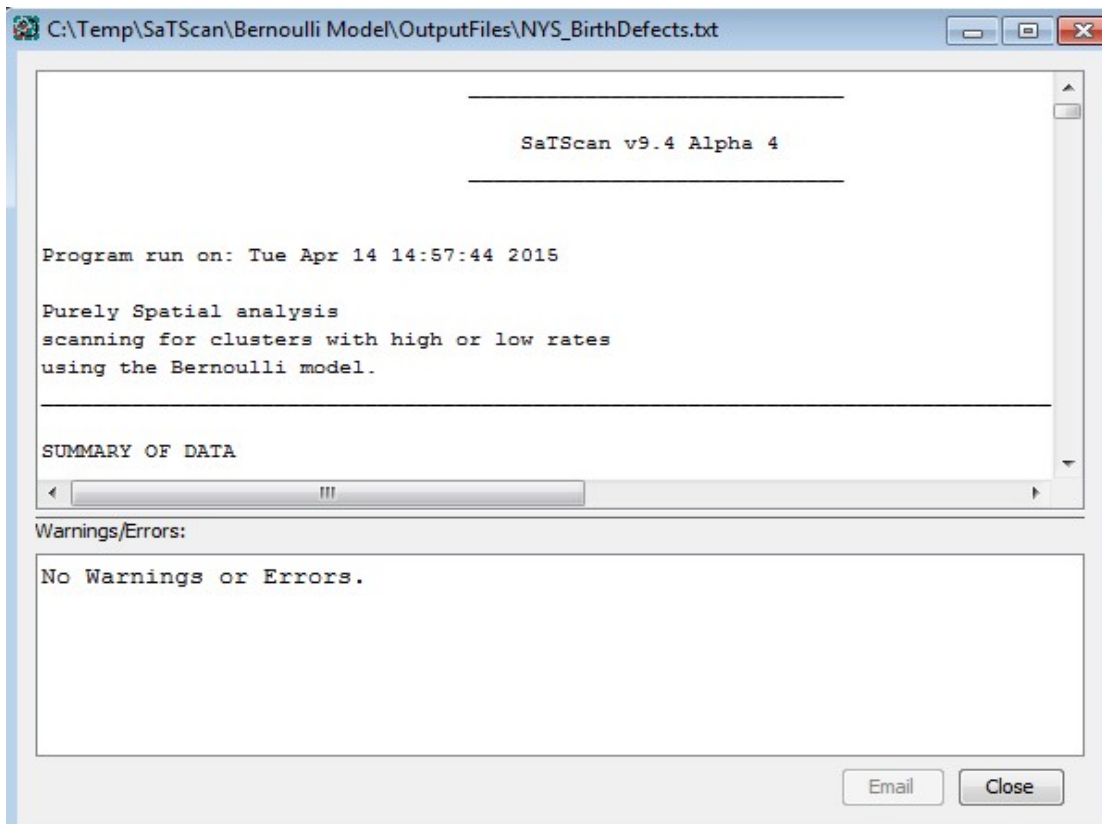
Sometimes SaTScan produces warnings or error. The most common errors are problems with the input data, such as a location ID that is present in the case file but missing in the geographical coordinates file. The descriptions of the warnings and errors are meant to help find problems that may exist in the input data. In this tutorial, you should not get any warnings or errors if you have done everything according to the tutorial instructions.

9. Reading and Interpreting the Results

The main text-based results file will automatically open once the calculations are 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 or low rates using the Bernoulli 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 is the right data. If all went well in this tutorial, it should look like this:

SUMMARY OF DATA
Study period........................: 2005/1/1 to 2009/12/31
Number of locations.................: 1143
Total population....................: 1237189
Total number of cases...............: 24940
Percent cases in area...............: 2.0

This means that in the 1143 zip-codes in New York State, there were a total of 1,237,189 births from 2005 to 2009. Of these, 2.0% had a birth defect, for a total of 24,940 children.

The clusters detected comes next, and this is the most important part of the results file. SaTScan will list the most likely cluster, that is, the cluster that is least likely to have occurred by chance, as well as a list secondary clusters. For the New York State birth defect data, there were seven statistically significant clusters. Of these, clusters #1, #2, #4 and #7 had more birth defects than expected, while clusters #3, #5 and #6 had fewer birth defects than expected. For each cluster, a detailed description is provided:

2. Location IDs included: GATid182, 14701, 14747, 14750
 Coordinates / radius..: (42.150829 N, 79.194379 W) / 12.36 km
 Population............: 3172
 Number of cases.......: 141
 Expected cases........: 63.94

Observed / expected...: 2.21
Relative risk.........: 2.21
Percent cases in area: 4.4
Log likelihood ratio..: 35.526304
P-value...............: 0.0000000000035

In this, the second most likely cluster, there were a total of 3172 birth in New York State, of which 141 births had a birth defect. This is 141/3172=4.4% of all births, compared to (24940-141)/(1237189-3172)=2.0% outside the cluster, corresponding to a relative risk of [141/3172] / [(24940-141)/(1237189-3172)] = 2.21. Assuming that this cluster had the same proportion of birth defects as the average in the state, the expected number of cases would have been 3172 * 24940 / 1237189 = 63.94. The observed / expected is then calculated as 141 / 63.94 = 2.21. Note that, for small clusters like this one, the relative risk is very close to the observed / expected, but for clusters with a large population, that is typically not the case. The relative risk compares what is inside the cluster to what is outside the cluster, while the observed / expected compares what is inside the cluster to what would have been inside the cluster had the null hypothesis been true.
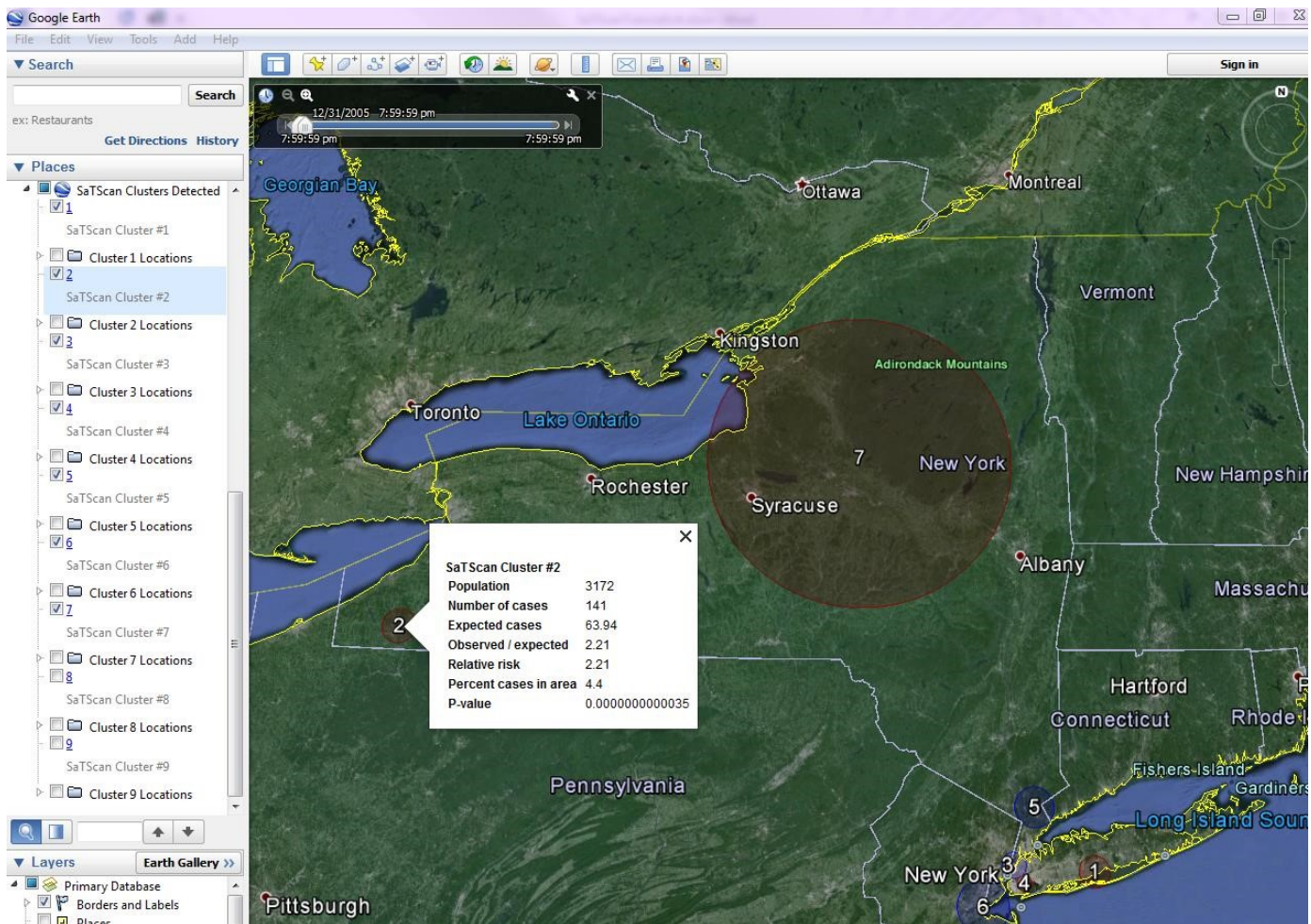
In the text based results file, you can determine the location of the circular cluster in two ways. One way is to look at the location IDs included in the cluster. A second way it to look at the latitude/longitude coordinates of the cluster centroid, together with the radius of the circle. An easier way is to draw the cluster on a map. Unless you request otherwise, SaTScan will automatically open the KML file if you have Google Earth installed on your computer. Alternatively, SaTScan will create either a .kml or a .shp shape file that you can use as an input file to the GIS software of your liking.

```
C:\Temp\SaTScan\Bernoulli Model\OutputFiles\NYS_BirthDefects.txt                          ─ □

Inference
---------

  Number of Replications              : 999

Spatial Output
--------------
  Automatically Launch Google Earth         : Yes
  Compress KML File into KMZ File           : No
  Include All Location IDs in the Clusters  : Yes
  Cluster Location Threshold - Separate KML : 1000
  Report Hierarchical Clusters              : Yes
  Criteria for Reporting Secondary Clusters : No Geographical Overlap
  Report Gini Optimized Cluster Collection  : No
  Restrict Reporting to Smaller Clusters    : Yes
  Reported Clusters                         : Only clusters smaller than 25 percent of population at risk reported.

Other Output
------------
  Report Critical Values   : No
  Report Monte Carlo Rank  : No
  Print ASCII Column Headers : Yes

Run Options
-----------
  Processer Usage    : All Available Proccessors
  Suppress Warnings  : No
  Logging Analysis   : Yes

_____

Program completed  : Tue Apr 14 14:57:49 2015
Total Running Time : 5 seconds
Processor Usage    : 8 processors
```
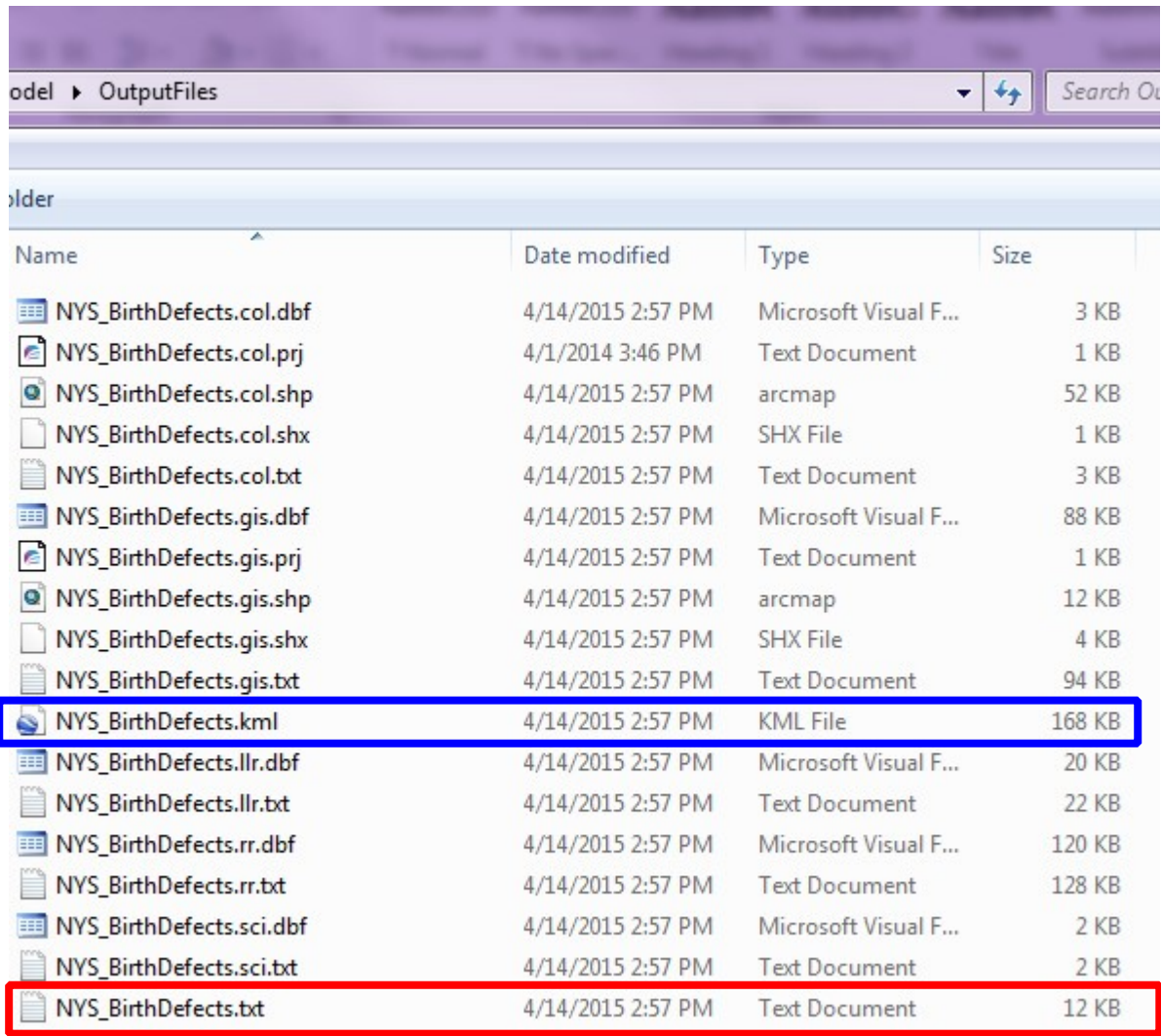
At the end of the main results file, there is a list of the parameter settings used for the analysis. For example, it states how many Monte Carlo replications were used. In this way, you can always go back and check the parameters you used for a particular analysis. At the very end, there is also information about when the SaTScan program was run, and how long it took to run it.

The list of parameter settings also includes the names of the optional results files that you requested, if any. These optional files contains the results in column and row format, for easy incorporation and further analysis using GIS and other software products. They are located in the same directory that you specified for the text based Results File, with the same

file name but different extensions. Nobody needs all these 18 output files, but what you need depends on which other software you are using. The two most commonly used output files are marked in the figure below, with the text based Results File indicated in red while the Google Earth .kml file is shown in blue.

10. References and Further Reading

This is the second in a series of SaTScan tutorials. As a complement to this one, we also recommend doing the first one, which uses the purely spatial Poisson model for cancer incidence data. Subsequent tutorials will cover space-time scan statistics; different probability models and different types of diseases.

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 web site: http://www.satscan.org/techdoc.html.

For scientific publications describing the Bernoulli based purely spatial scan statistic, we recommend:

General Statistical Theory, Bernoulli Probability Model

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

SaTScan Applications to Birth Defects Data

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.

Boyle E, Johnson H, Kelly A, McDonnell R. Congenital anomalies and proximity to landfill sites. Irish Medical Journal, 97:16-18, 2004.

Additional references of both a methodological and applied nature can be found in the SaTScan User Guide: http://www.satscan.org/techdoc.html.

Please contact the authors with any comments, questions or suggestions:

Thomas Talbot; Center for Environmental Health, New York State Department of Health
Thomas.Talbot@health.ny.gov

Sanjaya Kumar; Center for Environmental Health, New York State Department of Health
Sanjaya.Kumar@health.ny.gov

NOTES