

# SaTScan Tutorial #3

## Advanced Options

Abdurrahman Abdurrob  
Martin Kulldorff  
Brigham and Women's Hospital  
Harvard Medical School  
August, 2016

# Contents

Chapter One: Introduction.....	3
Chapter Two: Scan for High or Low Rates Only .....	5
Chapter Three: Geographical Subset Analyses.....	14
Chapter Four: P-Value and Monte Carlo Replications.....	36
Chapter Five: Maximum Cluster Size to Evaluate.....	44
Chapter Six: Spatial Clusters to Report .....	59
Chapter Seven: Gini Clusters.....	67
References and Further Reading .....	72
NOTES.....	73

# Chapter One: Introduction

## 1.1 Background

This is a step-by-step tutorial that highlights many of the advanced options and features in the SaTScan™ software. 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 clusters, and to determine if they are statistically significant, adjusting for the multiple testing inherent in the many possible cluster locations and sizes. The software was designed for disease surveillance but may also be used for similar problems in other fields such as archeology, criminology, demography, ecology, geography or zoology. A list of papers published in many different application areas can be found in the SaTScan bibliography: <http://www.satscan.org/references.html>

This third SaTScan tutorial uses the same data as SaTScan Tutorial #1, and the general goal is also the same, to use the purely spatial scan statistic to analyze the geographical distribution of female breast cancer incidence in New York State, in order to determine if there are any geographical clusters of breast cancer incidence. That is, we will determine if there are any geographical areas with more breast cancer cases than would be expected if the risk of breast cancer was evenly distributed across the State adjusted for age. The same purely spatial Poisson model will be used but we will describe and explore a few of the advanced features available in SaTScan. While we illustrate these advanced features using the Poisson model, most of them are also available for the other probability models in SaTScan.

## 1.2 New York State Breast Cancer Incidence Data

The data for this tutorial consists of female breast cancer incidence in New York State, for the years 2005 to 2009. The data comes from the New York State Cancer Registry. It can be downloaded from either the [New York State Department of Health website](#) or the [SaTScan web site](#). A detailed description of the data set is provided in [SaTScan Tutorial #1](#). In brief, the data set contains 72,296 observed breast cancer cases in 13,848 Department of Health (DOH) regions, and the age adjusted expected counts for those same regions.

## 1.3 Prerequisites

This tutorial is intended for self-learning, but it can also be used in a classroom setting. The prerequisite knowledge for this tutorial is a basic understanding of statistics and epidemiology. Before doing this third SaTScan tutorial, it is necessary to complete [SaTScan Tutorial #1](#), and to save the parameter file from that tutorial. How to do so is described

below. This tutorial is independent of SaTScan Tutorial #2, which does not have to be completed before this one.

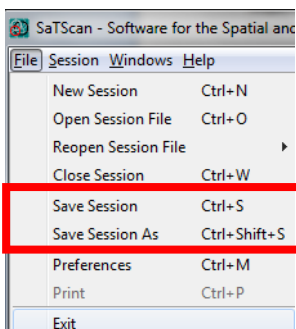
After completing this introductory chapter, chapters 2, 3, 4 and 5 can be read independently of each other, and in any order. Chapters 3 and 4 and Chapters 5, 6 and 7 are best read in that sequence. We recommend using the SaTScan User Guide as a complement to this tutorial, as it contains additional variants of the advanced features that will be covered.

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.

### 1.4 Save Parameter File from Tutorial #1

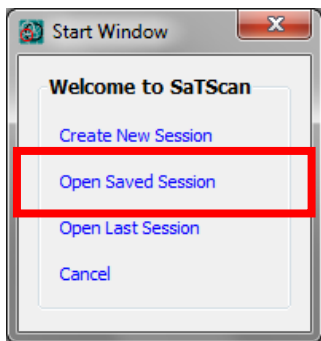
Doing SaTScan Tutorial #1 is a prerequisite for this chapter. That tutorial can be found on the [SaTScan web site](#). Once you have run that tutorial, please save the parameter file from Tutorial #1, as those settings will be used as the starting point for the analyses in this tutorial. This is done as follows:

To save the current SaTScan session with all its parameter settings, click on *'File'* and then *'Save Session'* or *'Save Session As'*.



### 1.5 Loading Parameter File

To open this Saved Session, open SaTScan and simply click *'Open Saved Session'* when prompted. Then locate the file name and directory, to which you saved the parameter file from Tutorial #1:



You are now ready to begin working through this tutorial.

## Chapter Two: Scan for High or Low Rates Only

### 2.1 Background

With the SaTScan software, it is possible to scan for areas with high rates of the disease, for areas with low rates of the disease, or simultaneously for areas with either high or low rates. The most common analysis is to scan for areas with high rates only. Sometimes though, the primary interest is to find areas with a low rate. For example, cases may be women who have received a mammography screening, with the goal of finding areas with low mammography screening rates. In other situations, there may be a simultaneous interest in finding both high and low rate areas.

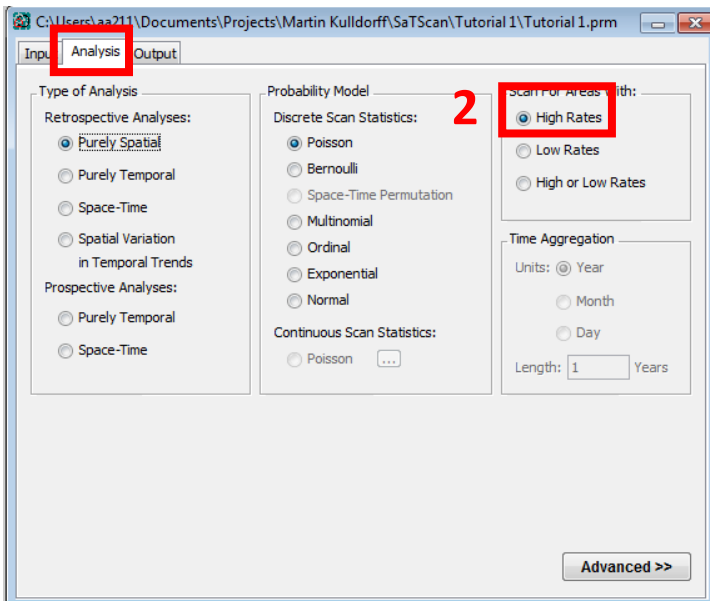
In a sense, the issue can be viewed as performing either a one-sided or two-sided statistical test. If there is only interest in high rate areas, one should only scan for high rate areas in order to maximize statistical power. The equivalent is true if one is only interested in low rate areas. It is important to note that running one analysis for both high and low rates should be used rather than running two separate tests for high rates and low rates respectively. The former will maintain the correct alpha level. With the latter approach, the null probability of having  $p < 0.05$  in at least one of the two analyses will be close to 0.10.

In SaTScan Tutorial #1, the New York State breast cancer data was analyzed using the purely spatial Poisson scan statistic, simultaneously scanning for both high and low rates. The goal of this chapter is to walk through the process of scanning for high rates only, comparing the results with those from Tutorial #1. For completeness, we will also analyze the data scanning for only low rates. While the practical public health importance of finding low rate clusters for breast cancer incidence is unclear, there are other data and research questions for which this feature can be very useful and applicable.

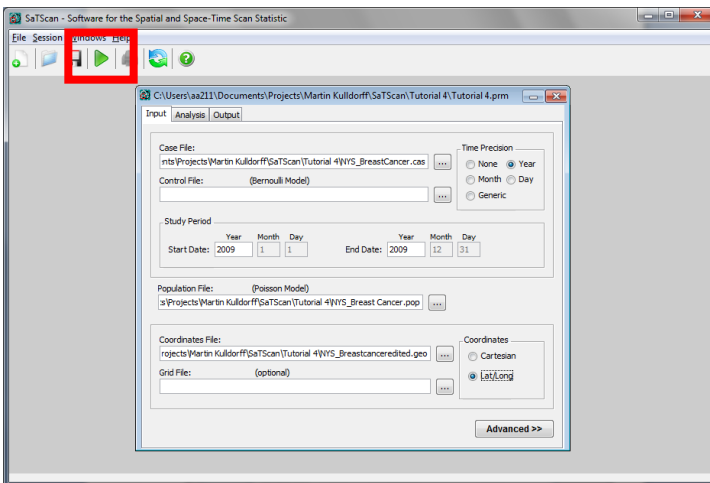
## 2.2 Scan for High Rates Only

First open the SaTScan session that was saved from Tutorial #1, as described in section 1.4 above. After loading the parameter file from Tutorial #1, switch over to the 'Analysis' tab highlighted below. The analysis in Tutorial #1 was done looking for clusters of 'High or Low Rates', but it will be run looking at high rates only by switching from 'High or Low Rates' to 'High Rates', as indicated below:

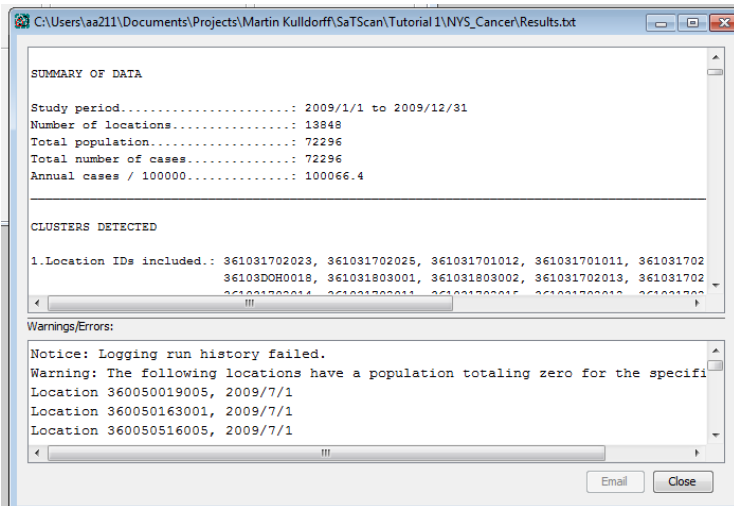
1



Run the analysis by hitting the green play button, highlighted below:

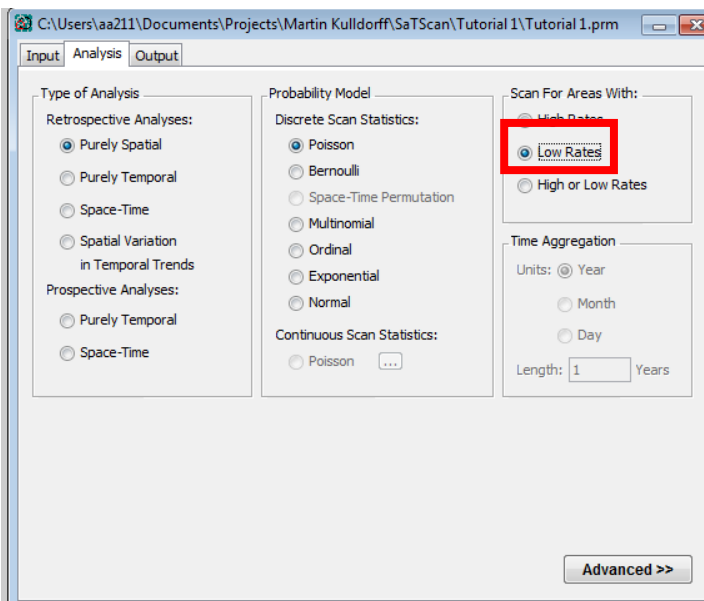


The following summary of data results will appear. In section 2.4, we will compare these results with those from Tutorial #1.

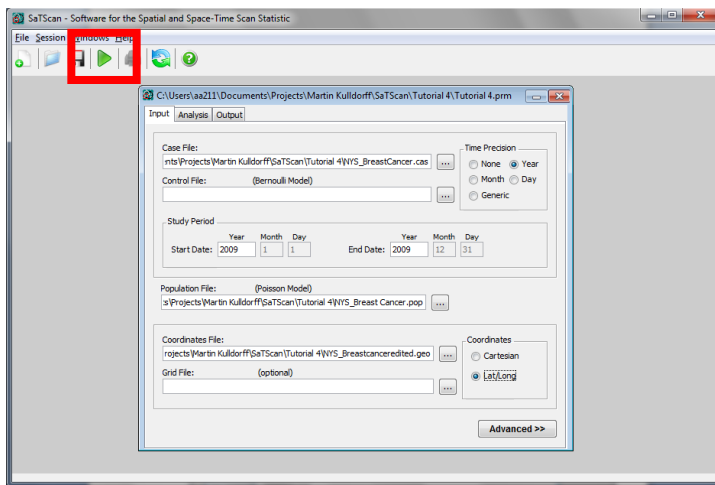


### 2.3 Scanning for Low Rates Only

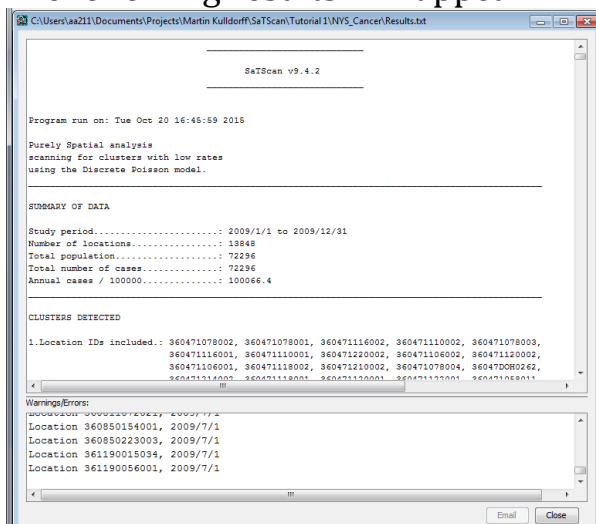
To scan for low rates only, switch from the current input for 'Scan for Area with' to 'Low Rates' on the main 'Analysis' Tab:



After that, run the analysis by hitting the green play button:



The following results will appear:



## 2.4 Comparison and Interpretation of Results

In this section, we compare the results for the high and low rate only analyses done above, with the results from the SaTScan Tutorial #1 results which simultaneously evaluated both high and low rates.

When Tutorial #1 results are rerun it is important to note that only significant clusters from the output will be displayed in Google Earth. For this comparison, we also want to look at some non-significant clusters. From the “High and Low Rates” analysis, we will manually add Cluster 5 to the Google Earth results. From the “Low Rates Only” analysis, we will also show Cluster 3 from the output.



To ensure cluster 5 is shown for the “High and Low Rates” analysis, please ensure that you have clicked the ‘5’ as shown below:

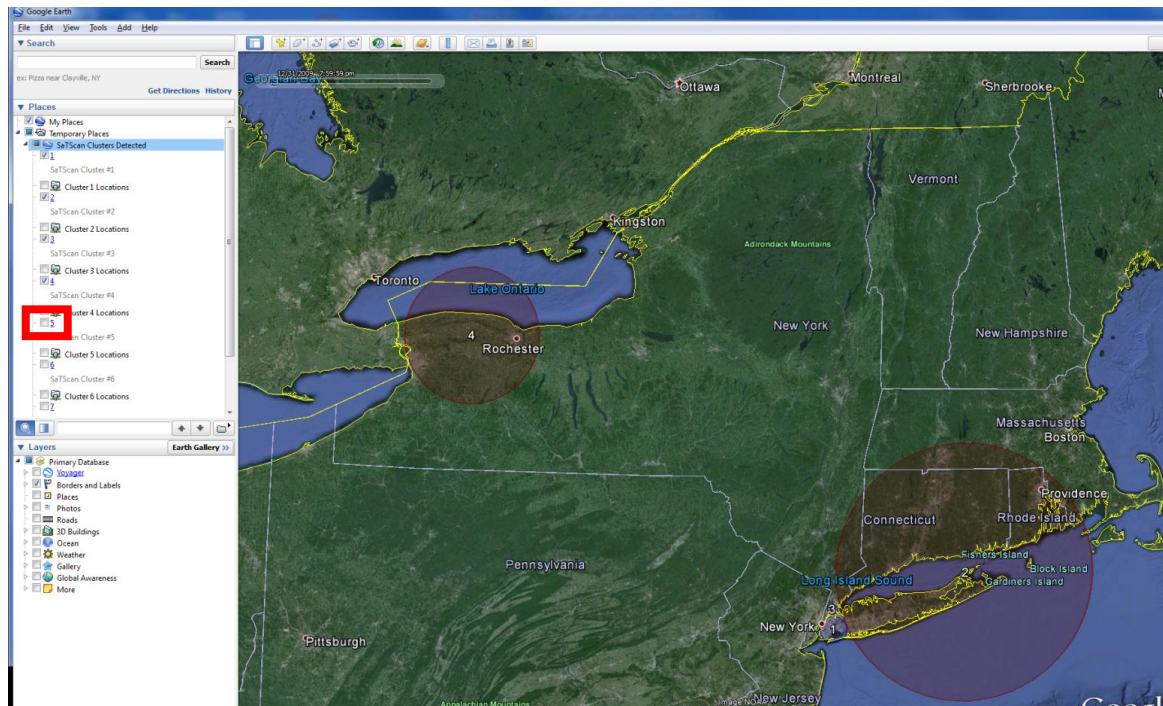


Figure 1: Tutorial #1 Output before Cluster 5

Please repeat this step for Cluster 3 from the output of the “Low Rates Only” analysis.

After doing so, cluster 5 will appear on the map as seen below:

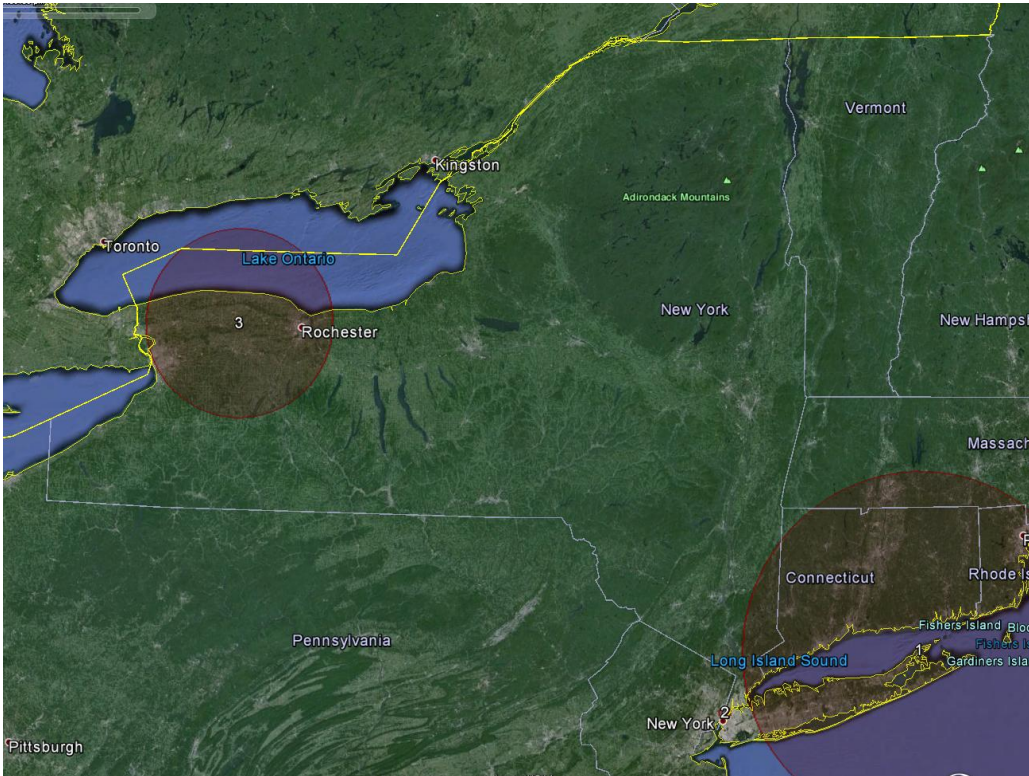


Figure 2: Results when scanning for high rates only

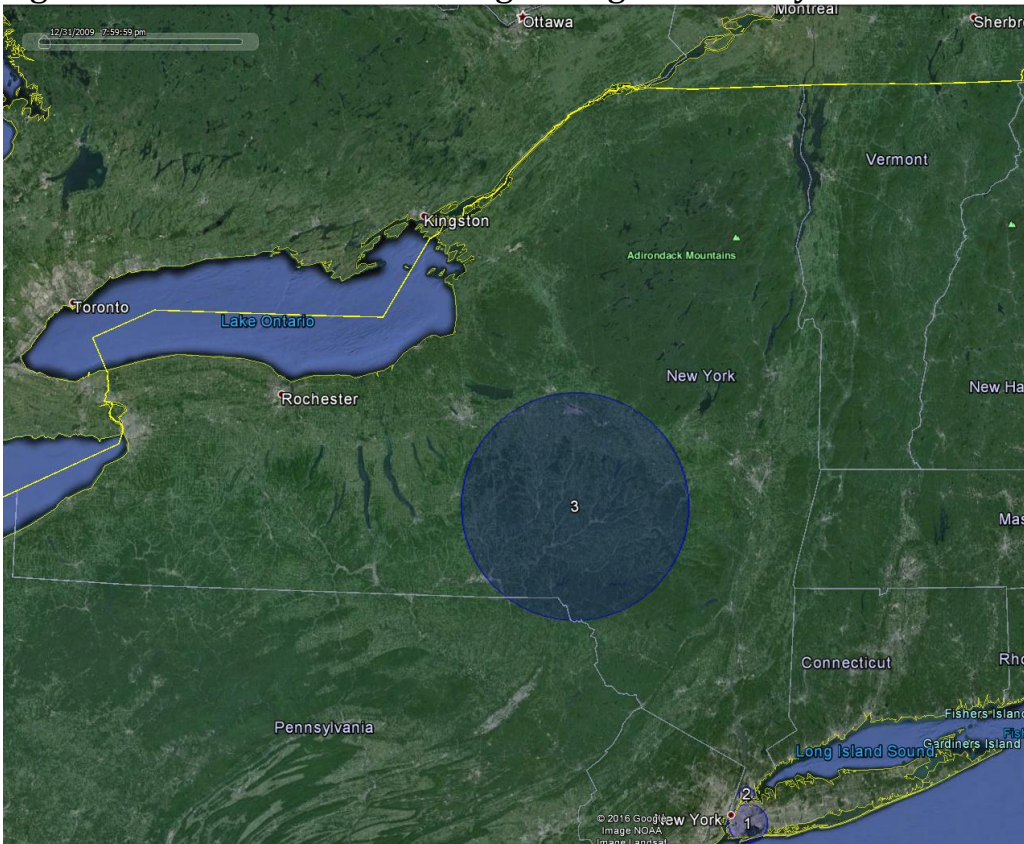


Figure 3: Results when scanning for low rates only showing all statistically significant clusters and one non-significant cluster

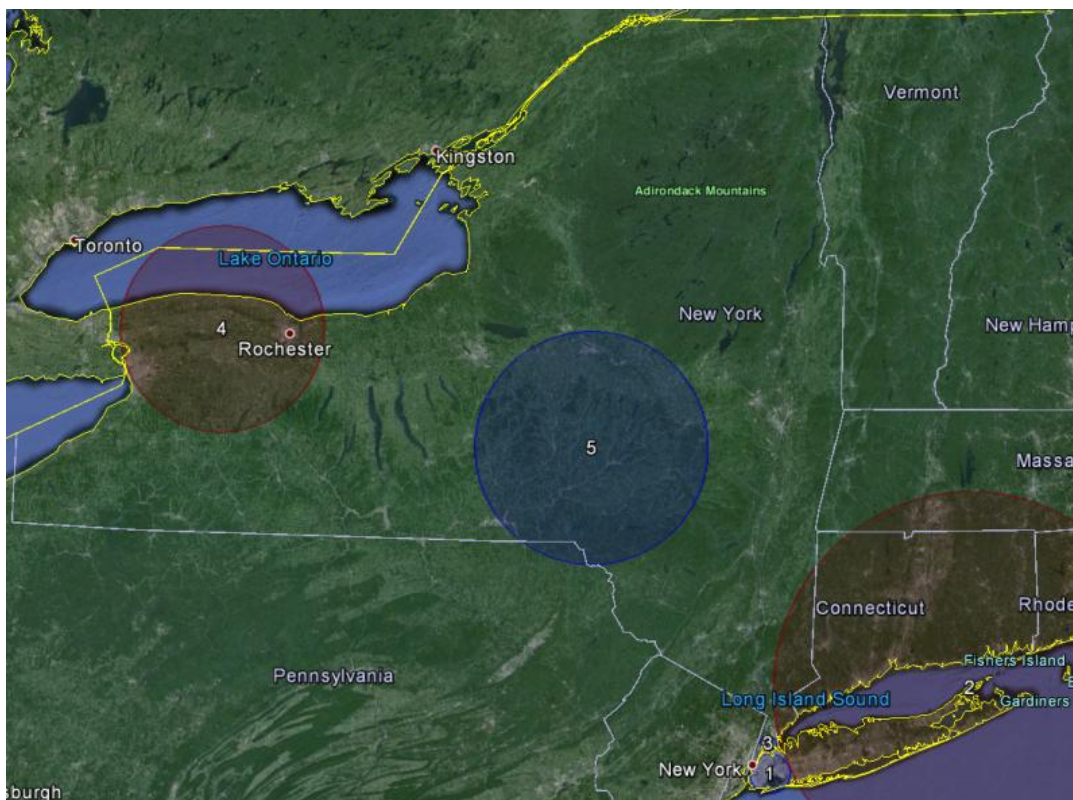


Figure 4: Results when scanning for high and low rates showing all statistically significant clusters and one non-significant cluster

<b>High Rates Only</b>	<b>Low Rates Only</b>	<b>Tutorial #1: High and Low Rates</b>
<b>Cluster 2</b> Coordinates / radius: (40.764710 N, 73.989910 W) / 4.08 km Observed Cases: 3648 Expected Cases: 2973.73 Relative risk: 1.24 P-value: < 0.0000001	<b>Cluster 1</b> Coordinates / radius: (40.659137 N, 73.873173 W) / 12.82 km Observed Cases: 13642 Expected Cases: 15886 Relative risk: 0.83 P-value: < 0.0000001	<b>Cluster 1</b> Coordinates / radius: (40.659137 N, 73.873173 W) / 12.82 km Observed Cases: 13642 Expected Cases: 15886 Relative risk: 0.83 P-value: < 0.0000001
<b>Cluster 1</b> Coordinates / radius: (41.126666 N, 72.339216 W) / 125.47 km Observed Cases: 15019 Expected Cases: 13416 Relative risk: 1.15 P-value: < 0.0000001	<b>No Overlapping Clusters</b>	<b>Cluster 2</b> Coordinates / radius: (41.126666 N, 72.339216 W) / 125.47 km Observed Cases: 15019 Expected Cases: 13416 Relative risk: 1.15 P-value: < 0.0000001
<b>No Overlapping Clusters</b>	<b>Cluster 2</b> Coordinates / radius: (40.835189 N, 73.884166 W) / 4.97 km Observed Cases: 3112 Expected Cases: 3976 Relative Risk.: 0.77 P-value: < 0.0000001	<b>Cluster 3</b> Coordinates / radius: (40.835189 N, 73.884166 W) / 4.97 km Observed Cases: 3112 Expected Cases: 3976 Relative risk: 0.77 P-value: < 0.0000001
<b>Cluster 3</b> Coordinates / radius: (43.174969 N, 78.154940 W) / 65.97 km Observed Cases: 7984 Expected Cases: 7098 Relative risk: 1.14 P-value: < 0.0000001	<b>No Overlapping Clusters</b>	<b>Cluster 4</b> Coordinates / radius: (43.174969 N, 78.154940 W) / 65.97 km Observed Cases.: 7984 Expected Cases: 7098 Relative risk: 1.14 P-value: < 0.0000001
<b>No Overlapping Clusters</b>	<b>Cluster 3</b> Coordinates / radius: (42.535144 N, 75.230508 W) / 74.03 km Observed Cases: 2010 Expected Cases: 2234 Relative risk: 0.90 P-value: 0.110	<b>Cluster 5</b> Coordinates / radius: (42.535144 N, 75.230508 W) / 74.03 km Observed Cases: 2010 Expected Cases: 2234 Relative risk: 0.90 P-value: 0.188

Table 1: Comparison of the High Rates Only and Low Rates Only with the Tutorial #1 results for both High and Low Rates. Geographically overlapping clusters are listed in the same row.

The clusters found in the high only analysis are similar to the high clusters found in the high and low analysis, and the clusters found in the low only analysis are identical to the low clusters found in the high and low analysis. For example, cluster #1 in the low only analysis is the same as cluster #1 in the high and low analysis. Cluster #1 in the high only analysis is the same as cluster #2 in the high and low analysis. The p-values may be different though. This can be seen when comparing cluster #3 in the low only analysis with cluster #5 in the high and low analysis. These clusters are identical, but cluster #3 in the low only analysis has a slightly lower p-value. This is because there is less multiple testing when only looking for low clusters, compared to looking for both low and high clusters.

As is natural, the analysis searching only for areas with high breast cancer incidence rates find clusters in very different locations than the analysis looking only for areas with low rates. Technically, it is possible to have some clusters that overlap, so that a location could belong to both a high rate and a low rate cluster. This is because not all location in a high rate cluster must have more cases than expected and vice versa. An example of this is seen in cluster 2 from the high only analysis which overlaps with cluster 1 from the high and low as well as the low only analysis.

The reason that this cluster does not show up in the results from the high and low analysis is that analysis had specified not to report overlapping clusters. If this requirement was loosened, to also reporting some overlapping clusters, these clusters would also show up in the analysis looking for either high or low rate areas.

## Chapter Three: Geographical Subset Analyses

### 3.1 Geographical Data Check

By default, SaTScan will check that all the cases and population numbers are at one of the locations specified in the geographical coordinates file. The reason is to make sure that there are no unintended data errors and to help the user to find any data problems that may exist.

It is possible to turn off this data checking procedure. Locations not in the geographical coordinate file are then ignored. This may be used if, for example, you only want to analyze a geographical subset of the data, in which case only the geographical coordinates file has to be modified while the other files can be used as they are. To explore this feature, we will examine breast cancer incidence in Upstate New York.

### 3.2 Breast Cancer Incidence in Upstate New York

In this chapter we evaluate the incidence of breast cancer in Upstate New York, by removing New York City, Long Island, Rockland County and Westchester County from the geographical coordinates file. That may detect high incidence clusters that are high compared to the rest of Upstate New York, but that was not high compared to New York City. This may also remove low incidence clusters that are low compared to New York City, but not significantly lower than the rest of Upstate New York.



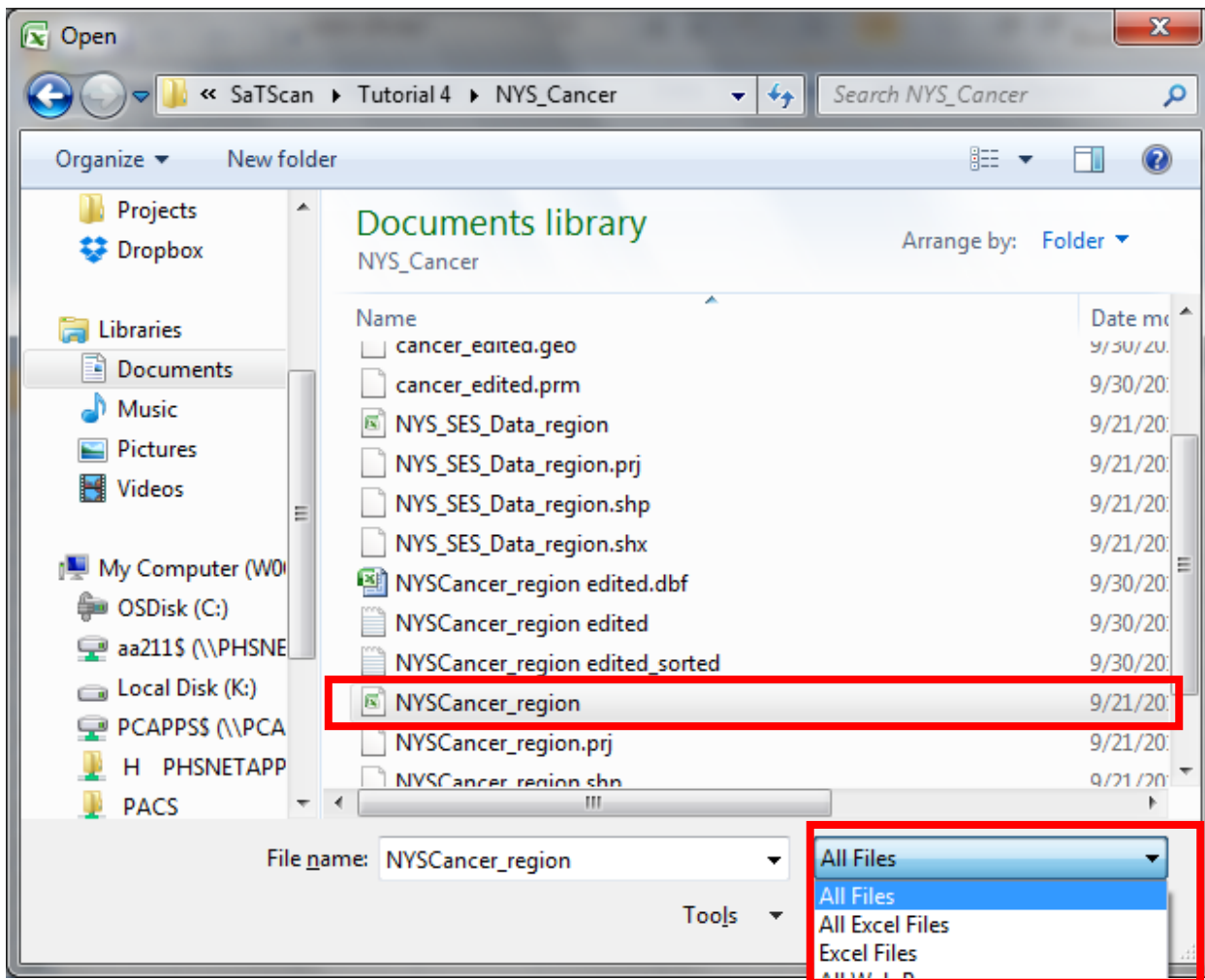
Figure 5: New York State with Upstate New York emphasized in green and the areas to be removed from the analysis in purple

The goal is for the user to quickly analyze a sub region by only having to edit the geographical coordinate file, rather than removing data from all of three input files. By not using the geographical data check, SaTScan will override the default data check,

proceeding with the analysis while ignoring those cases and population numbers that are not part of a location in the geographical coordinates file. Without this data check removed, SaTScan does not allow the analysis to be run, creating an error message.

### 3.3 Editing the Geographical Coordinates File

The first step is to edit the geographical coordinates file. Open Microsoft Excel or a similar editing program and then open the file NYS\_Cancer\_region.dbf. If you use Microsoft Excel, make sure to change the selection from Excel Files to All Files.



We will now show how to remove the data from New York City, Long Island and the southernmost part of the Hudson Valley. Each county in the United States has a FIPS code. A table is provided here for the FIPS codes for the specific counties in New York State that will be removed from this analysis:

County Name	FIPS Code
BRONX	36005
KINGS	36047
NASSAU	36059
NEW YORK	36061
QUEENS	36081
RICHMOND	36085
ROCKLAND	36087
SUFFOLK	36103
WESTCHESTER	36119

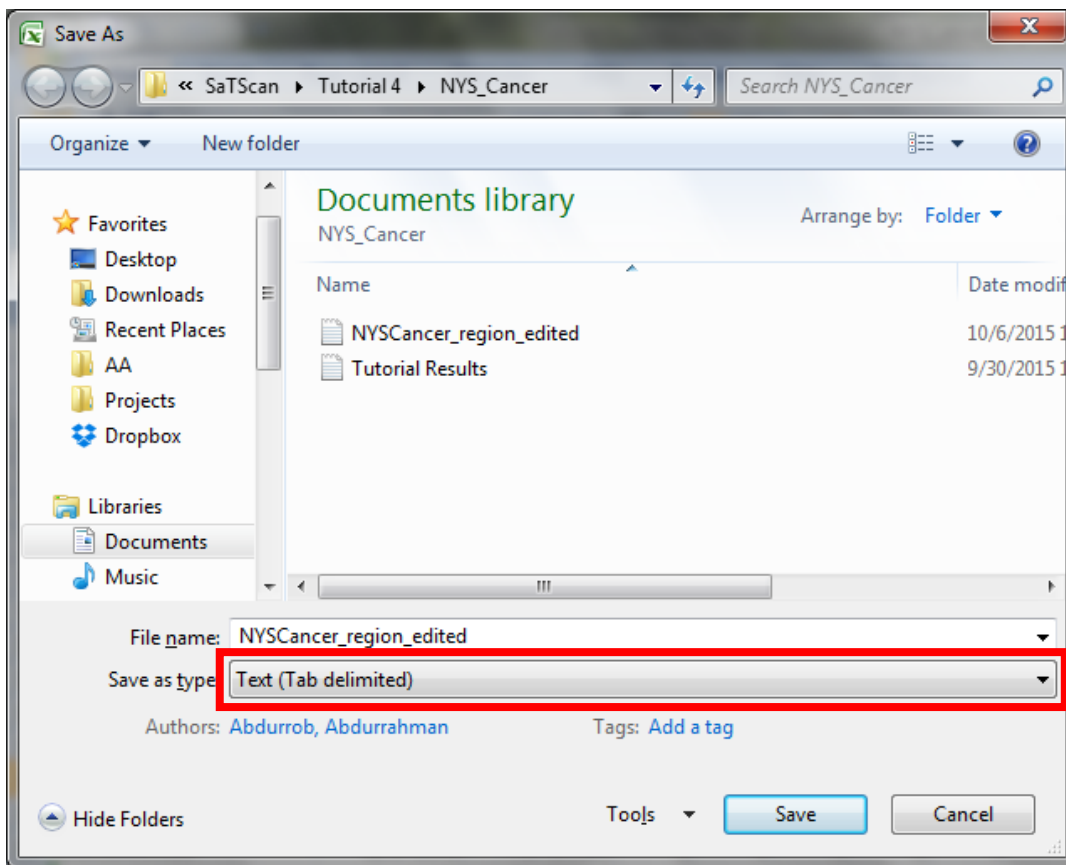
The first two digits of the FIPS code, '36', indicate that the county is in New York State. The following 3 digits are specific to the county. We will be matching these 5 digits to the first 5 digits of the 12 digit DOHREGION code in column A. For example, for Nassau the FIPS Code of 36059 corresponds to all the values in rows 5678 to 6790. By selecting them all and deleting those rows, they are removed from the file. Make sure that the actual rows are removed as it is not enough to just erase the content of the rows.

A5678		fx 360593001001						
	A	B	C	D	E	F	G	
1	DOHREGION	LATITUDE	LONGITUDE	OBLAD	OBONE	OBRAIN	OBREA	
5678	36059 001001	40.814276	-73.733211	3	0	0	7	
5679	36059 001002	40.814296	-73.720377	3	0	0	5	
5680	36059 001003	40.823548	-73.745522	1	1	0	4	
5681	36059 001004	40.816372	-73.757999	2	0	0	7	
5682	36059 003001	40.807313	-73.740400	7	0	0	6	
5683	36059 003002	40.798510	-73.737487	5	0	1	8	
5684	36059 003003	40.801714	-73.741263	1	0	1	4	
5685	36059 003004	40.809165	-73.743748	3	0	0	6	
5686	36059 004001	40.807241	-73.732128	1	0	1	14	
5687	36059 004002	40.802751	-73.728228	0	0	0	6	
5688	36059 004003	40.799279	-73.731790	6	0	2	6	
5689	36059 004004	40.798262	-73.717615	3	0	0	6	
5690	36059 004005	40.804242	-73.722542	1	0	0	7	
5691	360593005001	40.794797	-73.748495	0	0	0	5	

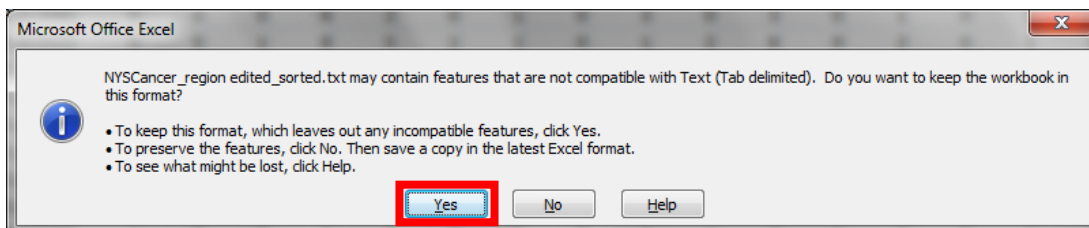
This process must be repeated for all nine counties listed above.

Once the database file has been successfully edited, save as a new file entitled: NYSCancer\_region\_edited. Save the file as a tab delimited text file. This will allow SaTScan to read the data file as distinct columns and rows.



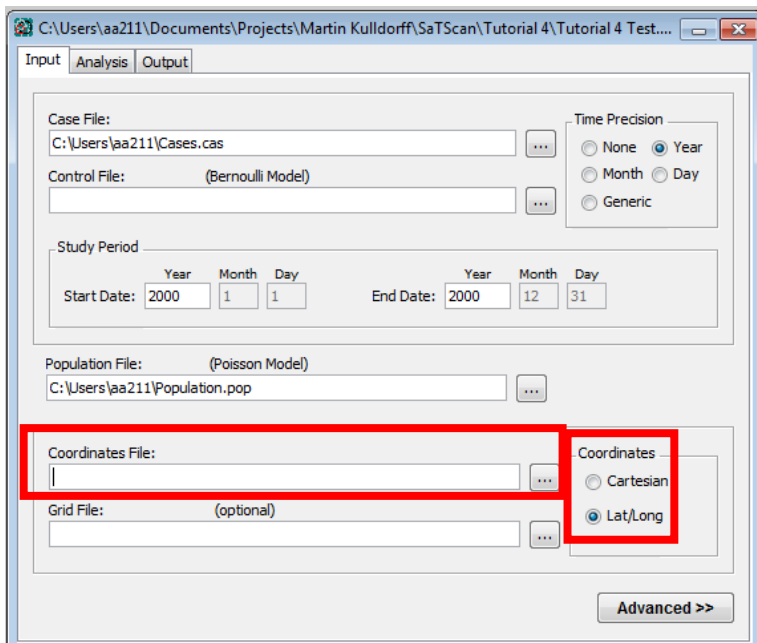


Make sure to select yes, to keep compatible:

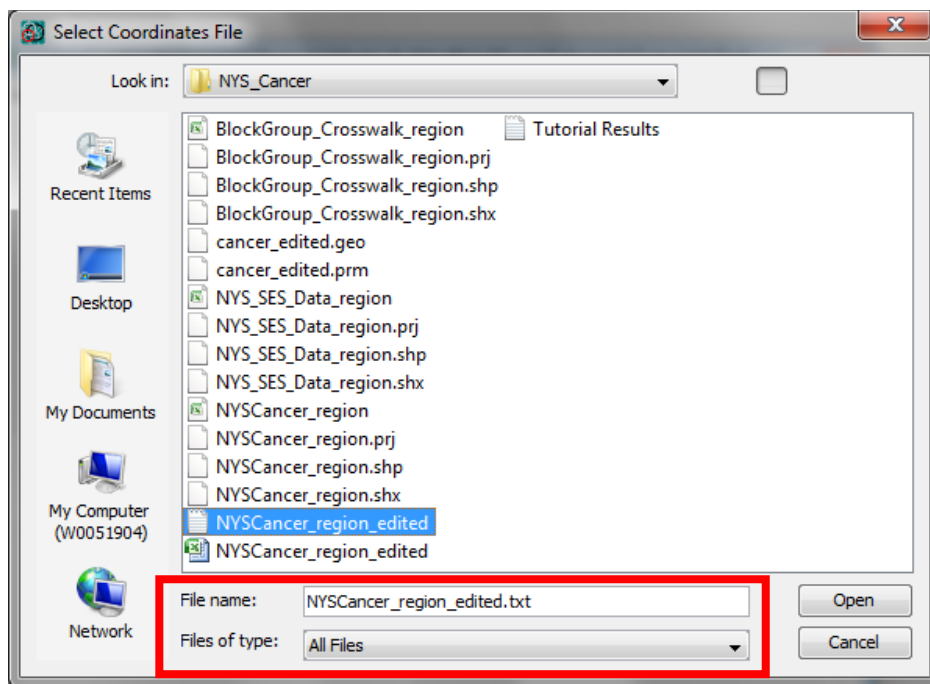


### 3.4 Running the Upstate New York Analysis

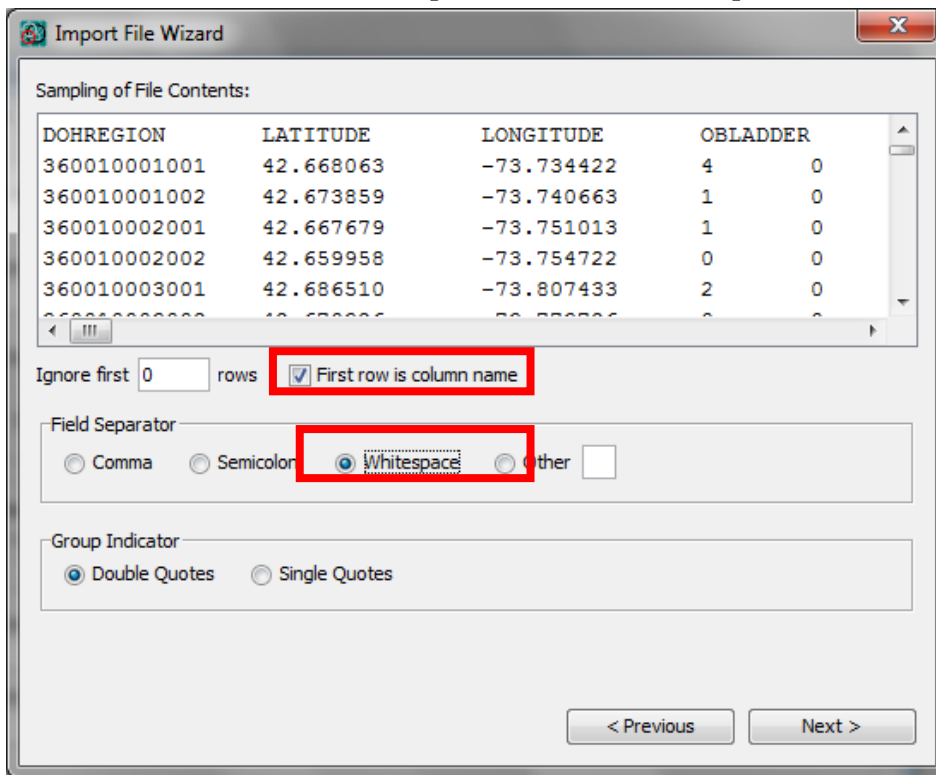
First open the SaTScan session that was saved from Tutorial #1, as described in section 1.4 earlier. For the Coordinate File, specify the use of the new file entitled: NYSCancer\_region\_edited.txt. It is also important to specify that this analysis should be done with the coordinate system of Lat/Long (highlighted below).



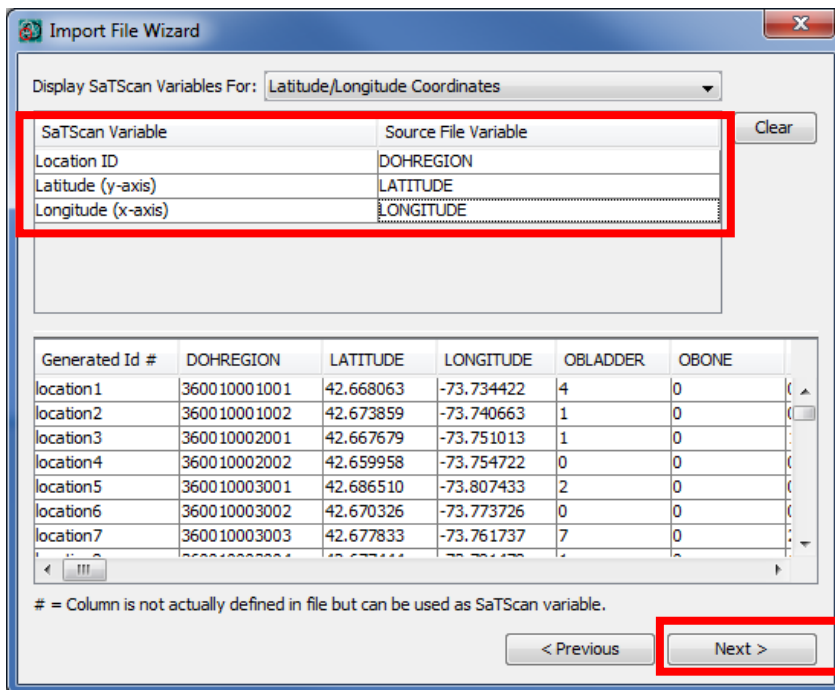
Again, import the file selecting 'All Files':



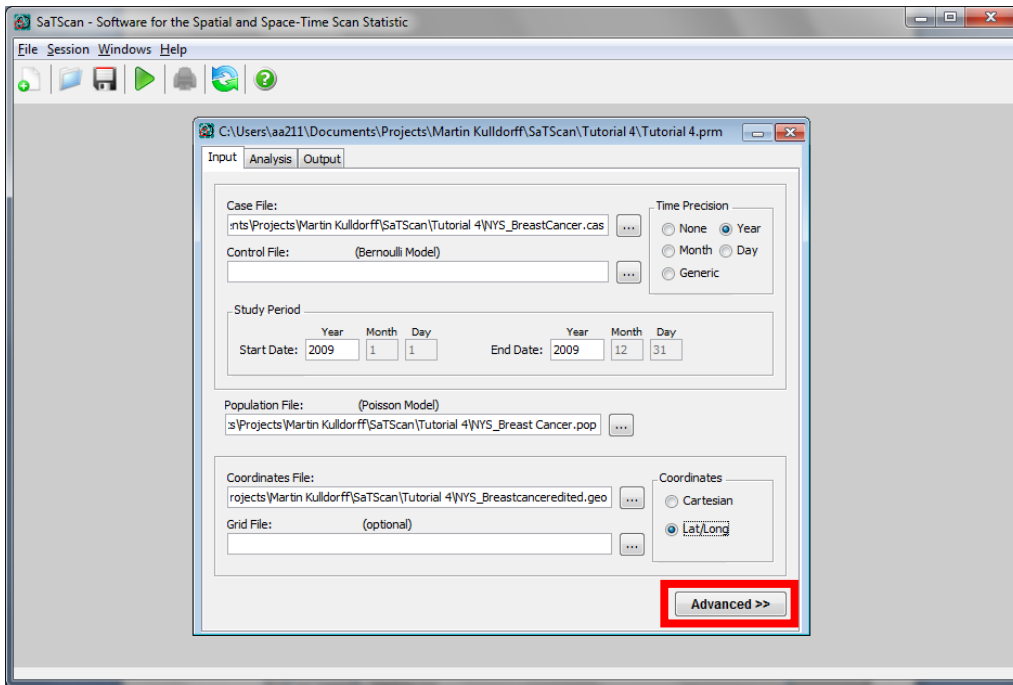
Make sure to select field separator as: 'Whitespace' and 'First row is column name'.



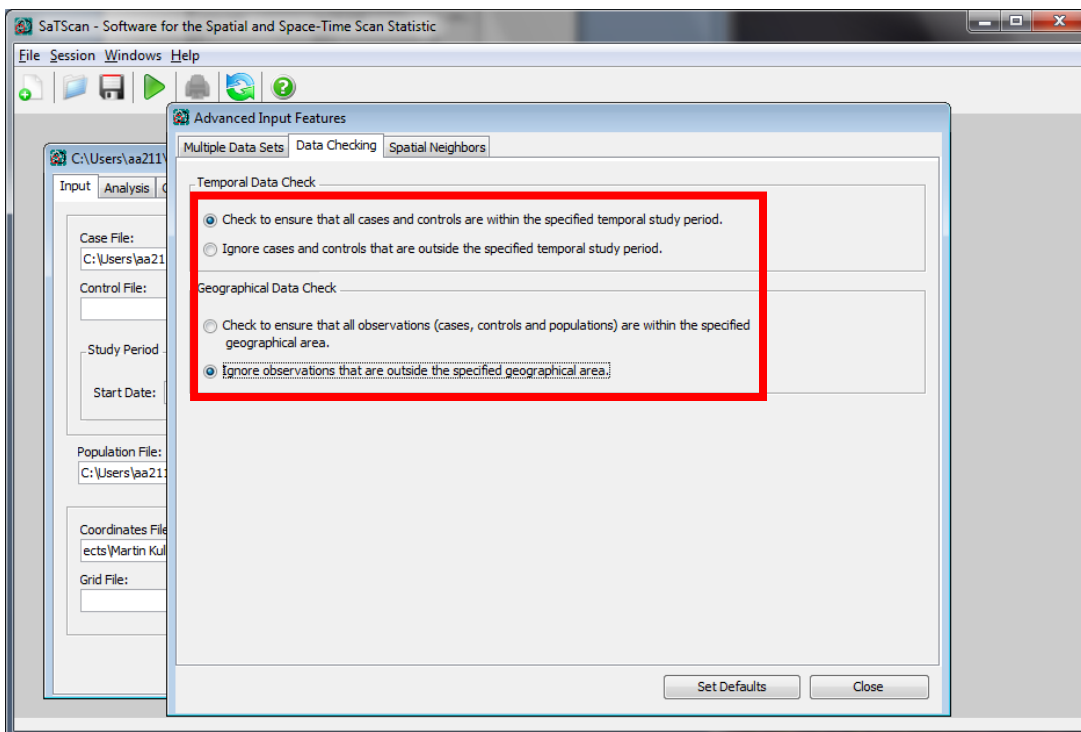
Now, select DOHREGION as the Location ID, LATITUDE as the Latitude (y-axis), and LONGITUDE as the Longitude (x-axis). Then proceed with next and follow the prompts until the Import File Wizard closes.



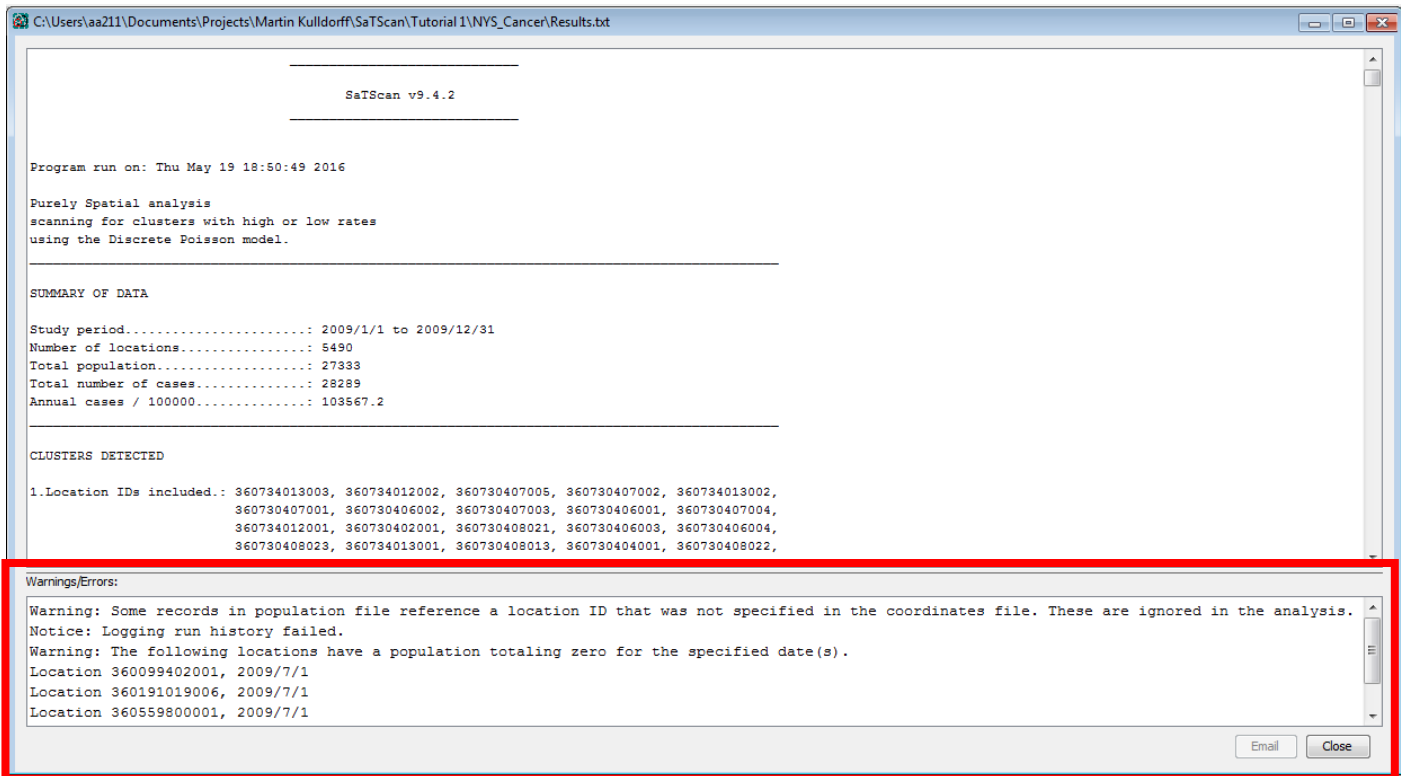
With the edited database uploaded, we must now deselect data checking in the advanced options on the 'Input' tab:



Deselecting the data check feature will allow you to run the analysis even though there are some locations listed in the case and population files that are not in the geographical coordinates file, and hence, not part of the analysis:



All other analysis parameters should be unchanged from Tutorial #1. Next, click on the green triangle to run the analysis. Note that there are a bunch of warnings in the Warning/Error section:



```
C:\Users\aa211\Documents\Projects\Martin Kulldorff\SaTScan\Tutorial 1\NYS_Cancer\Results.txt

SaTScan v9.4.2

Program run on: Thu May 19 18:50:49 2016

Purely Spatial analysis
scanning for clusters with high or low rates
using the Discrete Poisson model.

SUMMARY OF DATA

Study period.....: 2009/1/1 to 2009/12/31
Number of locations.....: 5490
Total population.....: 27333
Total number of cases.....: 28289
Annual cases / 100000.....: 103567.2

CLUSTERS DETECTED

1. Location IDs included.: 360734013003, 360734012002, 360730407005, 360730407002, 360734013002,
                           360730407001, 360730406002, 360730407003, 360730406001, 360730407004,
                           360734012001, 360730402001, 360730408021, 360730406003, 360730406004,
                           360730408023, 360734013001, 360730408013, 360730404001, 360730408022,

Warnings/Errors:
Warning: Some records in population file reference a location ID that was not specified in the coordinates file. These are ignored in the analysis.
Notice: Logging run history failed.
Warning: The following locations have a population totaling zero for the specified date(s).
Location 360099402001, 2009/7/1
Location 360191019006, 2009/7/1
Location 36059800001, 2009/7/1

Email Close
```

Since you deliberately excluded those locations, you can ignore these warnings.

### 3.5. Upstate New York Results

The results of the Upstate New York analysis is shown in Table 2, and compared with the Whole State analysis.

<b>Upstate New York Analysis</b>	<b>Whole State Analysis from Tutorial #1</b>
<b>No Match</b>	<b>Cluster 1</b> Coordinates / radius: (40.659137 N, 73.873173 W) / 12.82 km Observed Cases: 13642 Expected Cases: 15886 Relative risk: 0.83 P-value: < 0.0000001
<b>No Match</b>	<b>Cluster 2</b> Coordinates / radius: (41.126666 N, 72.339216 W) / 125.47 km Observed Cases: 15019 Expected Cases: 13416 Relative risk: 1.15 P-value: < 0.0000001
<b>No Match</b>	<b>Cluster 3</b> Coordinates / radius: (40.835189 N, 73.884166 W) / 4.97 km Observed Cases: 3112 Expected Cases: 3976 Relative risk: 0.77 P-value: < 0.0000001
<b>Cluster 1</b> Coordinates / radius: (43.174969 N, 78.154940 W) / 64.77 km Observed Cases: 7663 Expected Cases: 6820 Relative risk: 1.12 P-value: 0.00000000012	<b>Cluster 4</b> Coordinates / radius: (43.174969 N, 78.154940 W) / 65.97 km Observed Cases: 7984 Expected Cases: 7098 Relative risk: 1.14 P-value: < 0.0000001
<b>Cluster 2</b> Coordinates / radius: (42.535144 N, 75.230508 W) / 74.06 km Observed Cases: 2013 Expected Cases: 2237 Relative risk: 0.86 P-value: 0.0000048	<b>Cluster 5</b> Coordinates / radius: (42.535144 N, 75.230508 W) / 74.03 km Observed Cases: 2010 Expected Cases: 2234 Relative risk: 0.90 P-value: 0.188

Table 2: Comparison of the analysis for Upstate New York to All of New York

These clusters can be further visualized through Google Earth, which should have automatically opened. Cluster 5, which is not significant, has been visualized in Figure 6. For instructions on how to do this, see section 2.4.

## Upstate New York Analysis

## Tutorial #1: All of New York

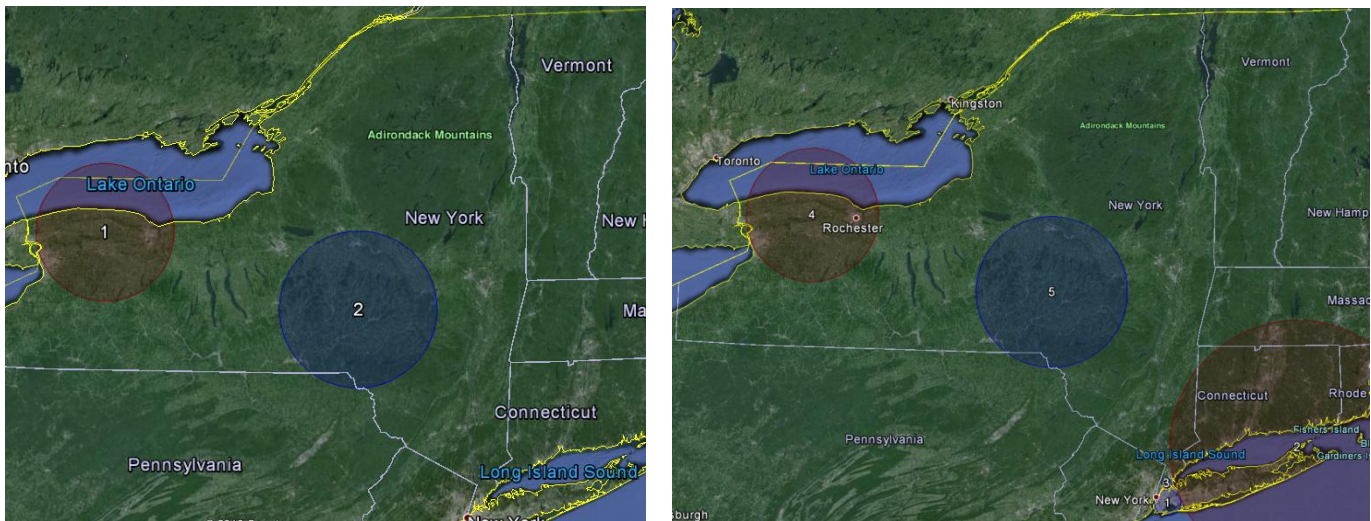


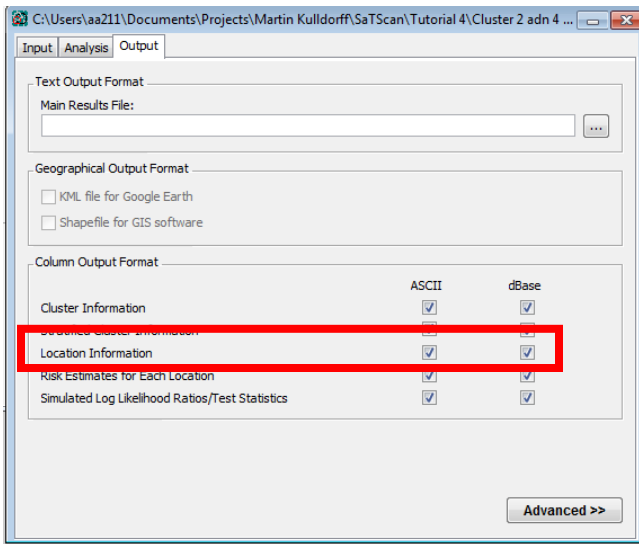
Figure 6: Maps of the results from the Upstate New York analysis (left) versus original Tutorial #1 analysis results for all of New York (right).

As seen in the new analysis, the same clusters in Upstate New York were found when excluding the New York City and metropolitan areas from the analysis specified earlier in the tutorial. The relative risks are slightly smaller though. This is because the breast cancer incidence is slightly higher in Upstate New York than in the New York City Metropolitan area.

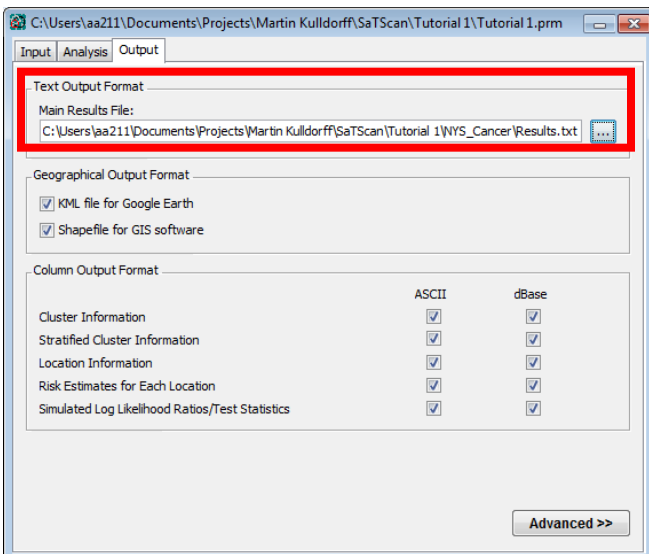
### 3.6. Looking for Clusters within a Cluster

When a detected cluster is large, it may be interesting to look for clusters within that cluster. For example, if we find a cluster with high rates, we may want to know if there are areas within that cluster that have exceptionally high rates. We will examine one high and one low cluster from the analysis done in Tutorial #1, for a deeper examination to see if there are significant differences in the incidence rates within those clusters. For this exercise, we selected the high incidence cluster near Buffalo (Cluster 4) and the low incidence cluster around Binghamton (Cluster 5).

Using the same editing concepts as previously described in Section 3.3, the DOH Regions for Clusters 4 and 5 from the High and Low Analysis in Tutorial #1 must first be isolated by recurring the analysis files from [Tutorial #1](#). Before rerunning that, select the 'Location Information' option in the 'Output' Tab is selected for both ASCII and dBase:

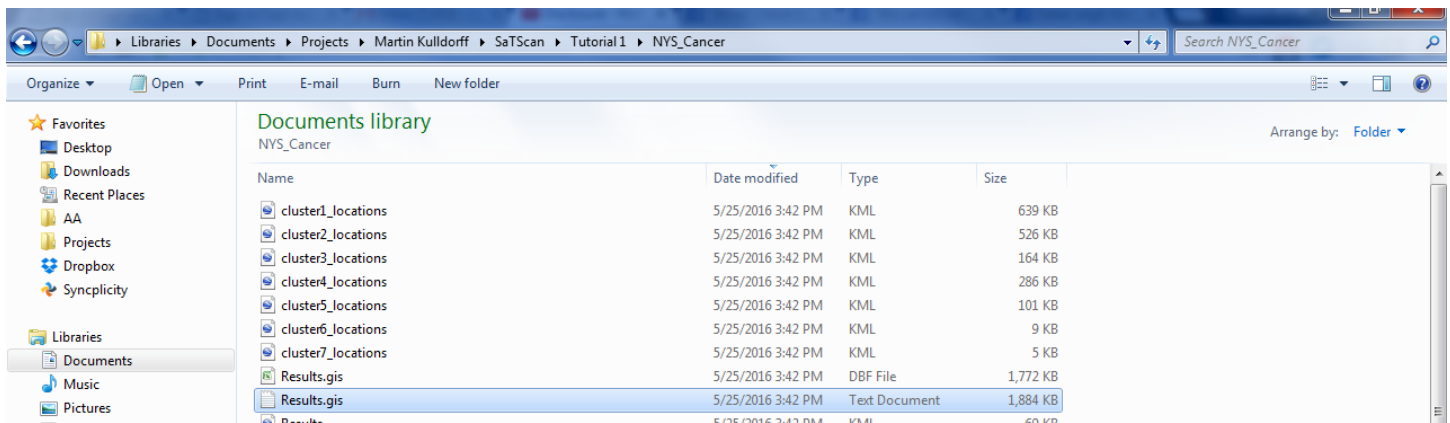


This is important because, the Location Information output will allow us to isolate the specific DOH Regions for Cluster 4 and 5. The Location Information results will output to a file with the same name as specified as the “*Text Output Format*”, but with the extension ‘gis’.

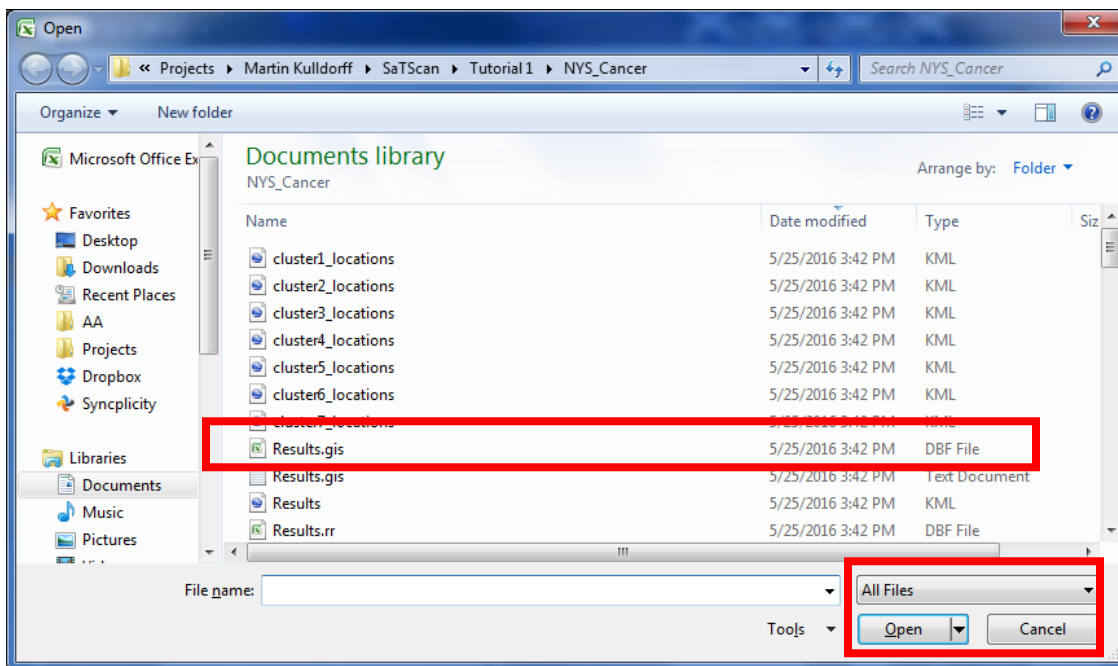


In the folder path specified above, the file titled “*Results.gis*” (below) contains the specific DOH regions that comprise each cluster. From this file, we will pull the DOH regions by opening the file in Microsoft Excel.





First, open and go to file-> open. Then proceed to the folder where the output is saved. Since the results file has a .gis extension, make sure to select “all files” when opening the file:



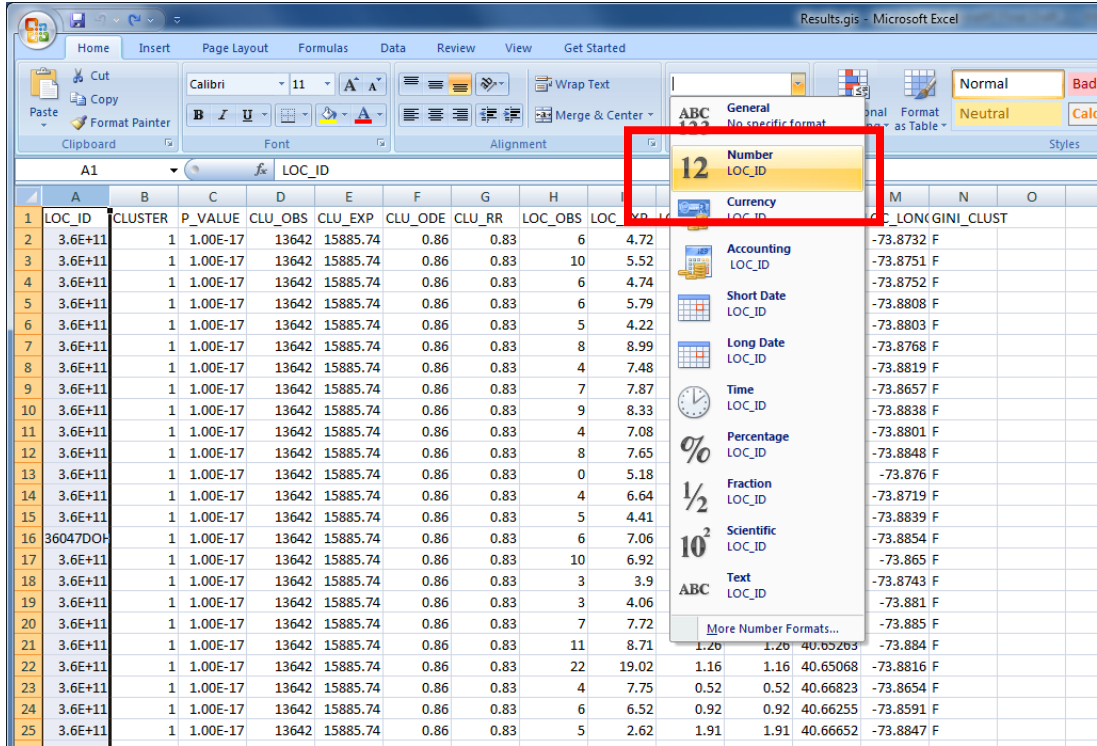
The DBF file (shown above) will import directly into excel correctly formatted, so make sure to choose this version and not the text file.

If imported correctly, the file should look like this:

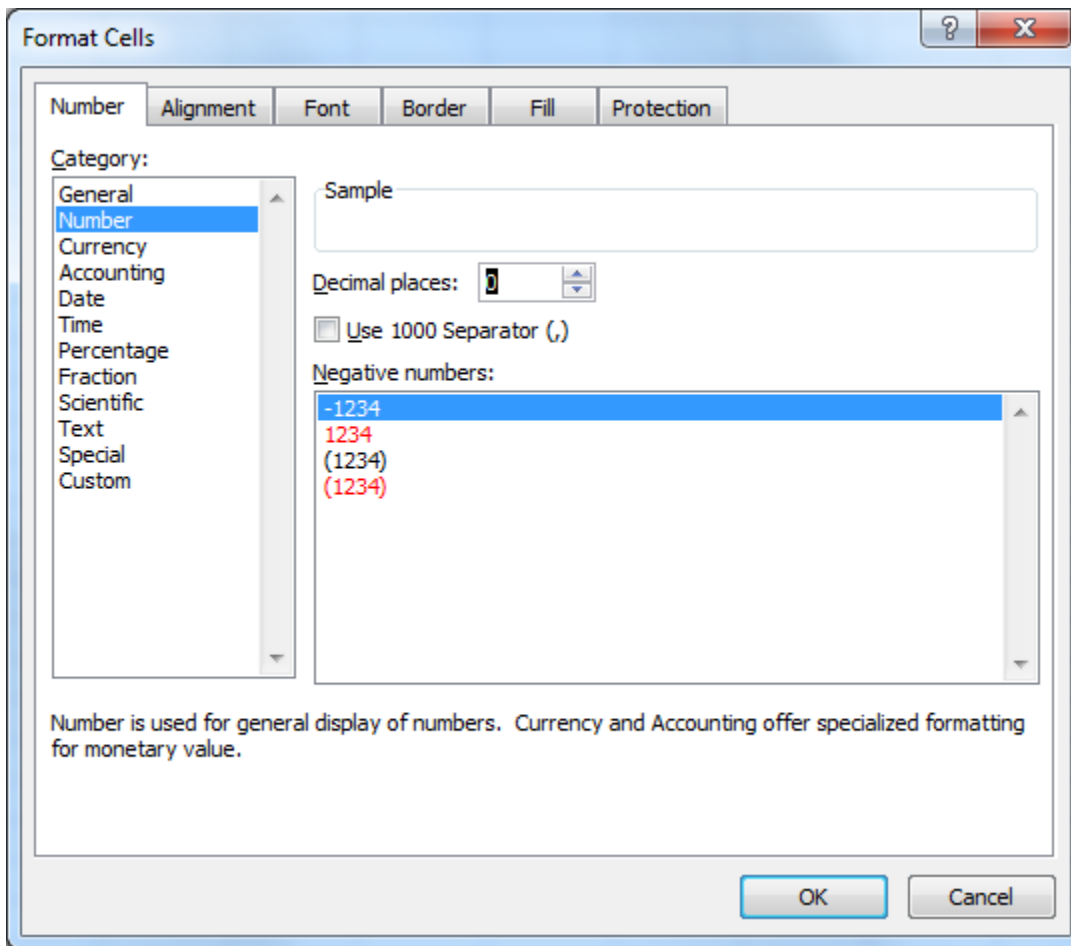
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
LOC_ID	CLUSTER	P_VALUE	CLU_OBS	CLU_EXP	CLU_ODE	CLU_RR	LOC_OBS	LOC_EXP	LOC_ODE	LOC_RR	LOC_LAT	LOC_LONG	GINI_CLUSTER							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	6	4.72	1.27	1.27	40.65914	-73.8732	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	10	5.52	1.81	1.81	40.66025	-73.8751	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	6	4.74	1.27	1.27	40.66315	-73.8752	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	6	5.79	1.04	1.04	40.65974	-73.8808	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	5	4.22	1.19	1.19	40.65608	-73.8803	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	8	8.99	0.89	0.89	40.66541	-73.8768	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	4	7.48	0.53	0.53	40.66128	-73.8819	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	7	7.87	0.89	0.89	40.66456	-73.8657	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	9	8.33	1.08	1.08	40.65858	-73.8838	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	4	7.08	0.57	0.57	40.66619	-73.8801	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	8	7.65	1.05	1.05	40.65799	-73.8848	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	0	5.18	0	0	40.66798	-73.876	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	4	6.64	0.6	0.6	40.6682	-73.8719	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	5	4.41	1.13	1.13	40.65483	-73.8839	F							
36047DOH	1	1.00E-17	13642	15885.74	0.86	0.83	6	7.06	0.85	0.85	40.66092	-73.8854	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	10	6.92	1.45	1.45	40.66663	-73.865	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	3	3.9	0.77	0.77	40.66926	-73.8743	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	3	4.06	0.74	0.74	40.66742	-73.881	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	7	7.72	0.91	0.91	40.66406	-73.885	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	11	8.71	1.26	1.26	40.65263	-73.884	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	22	19.02	1.16	1.16	40.65068	-73.8816	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	4	7.75	0.52	0.52	40.66823	-73.8654	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	6	6.52	0.92	0.92	40.66255	-73.8591	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	5	2.62	1.91	1.91	40.66652	-73.8847	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	3	6.79	0.44	0.44	40.66999	-73.8682	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	3	3.28	0.91	0.91	40.66911	-73.8808	F							
36047DOH	1	1.00E-17	13642	15885.74	0.86	0.83	17	13.96	1.22	1.22	40.64753	-73.8694	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	8	8.5	0.94	0.94	40.6711	-73.8726	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	1	3.62	0.28	0.28	40.66269	-73.8884	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	0	4.3	0	0	40.67129	-73.8767	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	2	3.58	0.56	0.56	40.66578	-73.8872	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	4	5.26	0.76	0.76	40.65868	-73.8564	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	9	5.25	1.71	1.71	40.67004	-73.8644	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	2	3.92	0.51	0.51	40.66221	-73.8901	F							
36047DOH	1	1.00E-17	13642	15885.74	0.86	0.83	8	8.5	0.94	0.94	40.65589	-73.8904	F							
3.6E+11	1	1.00E-17	13642	15885.74	0.86	0.83	14	11.85	1.18	1.18	40.65098	-73.8874	F							

This file contains Location ID in column A and both the cluster number in column B and the latitude and longitude in column L and M.

It is also important to convert Column A to “Number” Format otherwise SaTScan will not be able to differentiate the numbers that have the same scientific notation:



Finally, right click and select column A and go to “Format Call”. Here it is important to change the number of decimal places to 0 and then click OK.



### 3.7. Cluster 4 Sub Analysis

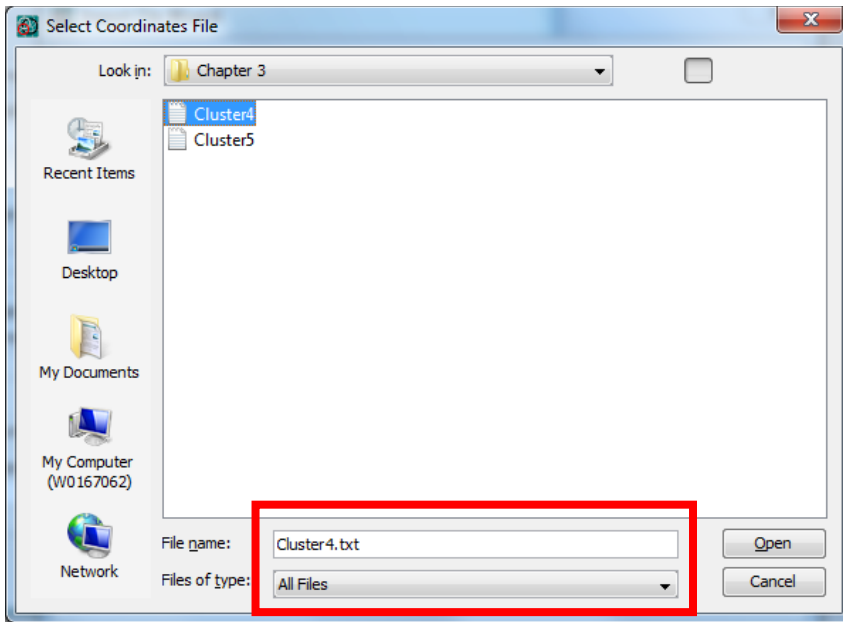
To create the coordinate file for the cluster 4 sub analysis, we will delete all entries that do not have a 4 in column B:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
6271	3.61E+11	3	1.00E-17	3112	3976	0.78	0.77	2	3.45	0.58	0.58	40.85024	-73.9391	F
6272	3.61E+11	3	1.00E-17	3112	3976	0.78	0.77	5	8.59	0.58	0.58	40.839	-73.9424	F
6273	3.61E+11	3	1.00E-17	3112	3976	0.78	0.77	3	3.54	0.85	0.85	40.85601	-73.9358	F
6274	3.61E+11	3	1.00E-17	3112	3976	0.78	0.77	1	4.79	0.21	0.21	40.8344	-73.9428	F
6275	3.6E+11	3	1.00E-17	3112	3976	0.78	0.77	0	2.67	0	0	40.84761	-73.8279	F
6276	3.61E+11	3	1.00E-17	3112	3976	0.78	0.77	5	6.87	0.73	0.73	40.8674	-73.9245	F
6277	3.6E+11	3	1.00E-17	3112	3976	0.78	0.77	2	3.28	0.61	0.61	40.82211	-73.8281	F
6278	3.6E+11	3	1.00E-17	3112	3976	0.78	0.77	1	3.57	0.28	0.28	40.87268	-73.8526	F
6279	3.61E+11	3	1.00E-17	3112	3976	0.78	0.77	4	6.57	0.61	0.61	40.82224	-73.9405	F
6280	3.61E+11	3	1.00E-17	3112	3976	0.78	0.77	5	6.14	0.81	0.81	40.84514	-73.9416	F
6281	3.61E+11	3	1.00E-17	3112	3976	0.78	0.77	7	4.02	1.74	1.74	40.84338	-73.9421	F
6282	3.61E+11	3	1.00E-17	3112	3976	0.78	0.77	3	5.1	0.59	0.59	40.84684	-73.9411	F
6283	36005DOH	3	1.00E-17	3112	3976	0.78	0.77	3	6.63	0.45	0.45	40.85153	-73.8292	F
6284				112	3976	0.78	0.77	4	7.23	0.55	0.55	40.84865	-73.9404	F
6285				112	3976	0.78	0.77	3	4.13	0.73	0.73	40.87527	-73.9103	F
6286				84	7098.18	1.12	1.14	4	2.85	1.4	1.4	43.17497	-78.1549	F
6287				84	7098.18	1.12	1.14	6	4.73	1.27	1.27	43.23306	-78.1802	F
6288				84	7098.18	1.12	1.14	10	4.34	2.3	2.3	43.17205	-78.2519	F
6289				84	7098.18	1.12	1.14	1	4.48	0.22	0.22	43.21112	-78.0711	F
6290				84	7098.18	1.12	1.14	4	4.58	0.87	0.87	43.24022	-78.1978	F
6291				84	7098.18	1.12	1.14	7	9.85	0.71	0.71	43.16513	-78.0563	F
6292				84	7098.18	1.12	1.14	3	2.52	1.19	1.19	43.25093	-78.1798	F
6293				84	7098.18	1.12	1.14	3	5.15	0.58	0.58	43.25308	-78.1549	F
6294				84	7098.18	1.12	1.14	6	5.07	1.18	1.18	43.25615	-78.1858	F
6295				84	7098.18	1.12	1.14	5	3.98	1.26	1.26	43.23525	-78.071	F
6296				84	7098.18	1.12	1.14	3	4.34	0.69	0.69	43.08829	-78.1274	F
6297	3.61E+11	4	1.00E-17	7984	7098.18	1.12	1.14	6	7.12	0.84	0.84	43.25341	-78.2232	F
6298	3.6E+11	4	1.00E-17	7984	7098.18	1.12	1.14	5	4.48	1.12	1.12	43.08329	-78.1775	F

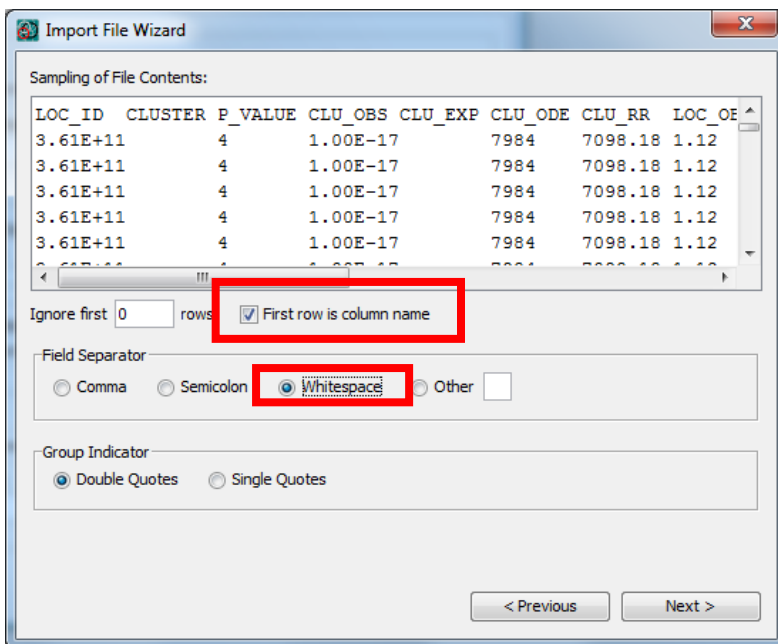
Once this is done, save this excel as a tab delimited text file called “Cluster 4.txt”. Instructions on how to save an excel file as a text file are in section 3.3.

Please open the saved parameter file for tutorial 1, as described in Section 1.5. These new files will be uploaded into SaTScan as the coordinate file through the Import Wizard.

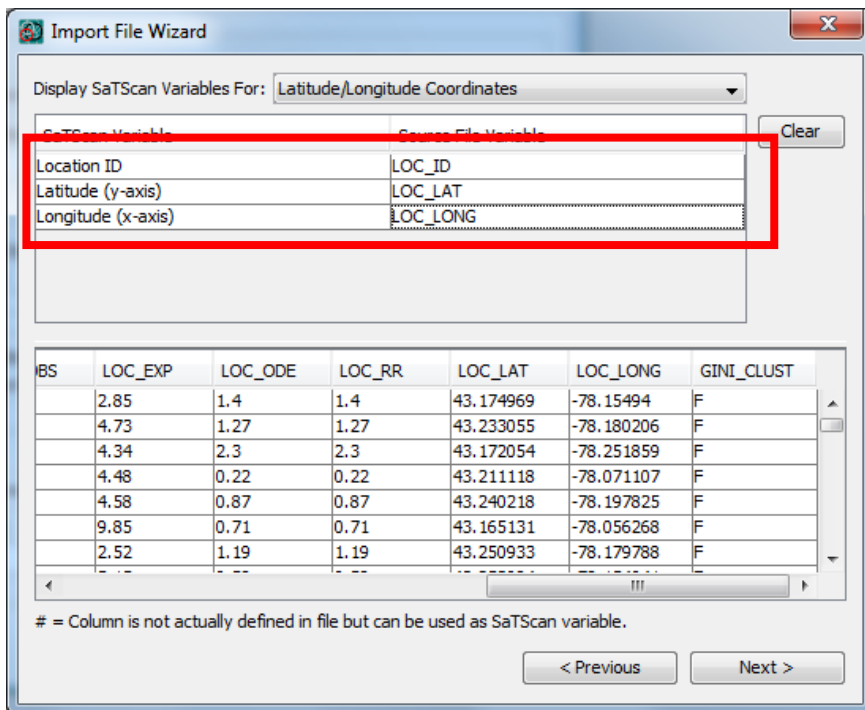
After opening the Import Wizard, make sure to select “All Files” for the “Files of Type”.



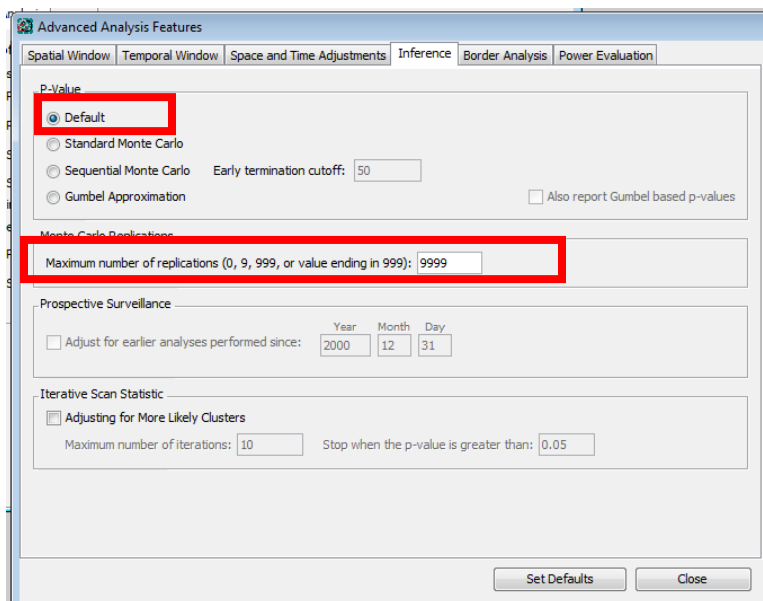
Make sure to select field separator as: 'Whitespace' and 'First row is column name'.



Make sure to select "LOC\_ID", "LOC\_LAT" and "LOC\_LONG" for Location ID, Latitude, and longitude respectively.



After following the remaining prompts, run the analysis with 9999 Replications in the Default setting for P-Value under the 'Inference' Tab:



Ensure that the geographical data check is off before running. This can be done by starting at the "Input" tab, going to the "Advanced" tab, "Data Checking" then finally clicking "Ignore observations that are outside the specified geographical area".

After running, the output indicates only one significant cluster as seen in Figure 7, which is shown below:

```

C:\Users\aa211\Documents\Projects\Martin Kuldorff\SaTScan\Tutorial 1\NYS_Cancer\Results.txt

SaTScan v9.4.2

Program run on: Thu May 19 19:00:14 2016

Purely Spatial analysis
scanning for clusters with high or low rates
using the Discrete Poisson model.

SUMMARY OF DATA
Study period.....: 2009/1/1 to 2009/12/31
Number of locations.....: 1349
Total population.....: 7098
Total number of cases.....: 7994
Annual cases / 100000.....: 112554.3

CLUSTERS DETECTED
1. Location IDs included.: 360550126006, 360550127001, 360550126002, 360550126004, 360550127004,
360550126001, 360550078021, 360550126003, 360550127005, 360550127006,
360550126005, 360550129001, 360550127003, 360550125002, 360550127002,
360550125004, 360550078022, 360550077001, 360550129002, 360550077002,
360550128002, 360550129003, 360550077003, 360550125003, 360550078011,
360550130021, 360550078012, 360550076001, 360550080064

Coordinates / radius...: (43.132141 N, 77.546365 W) / 2.70 km
Population.....: 115
Number of cases.....: 190
Expected cases.....: 129.67
Annual cases / 100000.: 164926.7
Observed / expected...: 1.47
Relative risk.....: 1.48
Log likelihood ratio...: 12.490330
P-value.....: 0.017

Warnings/Errors:
Warning: Some records in population file reference a location ID that was not specified in the coordinates file. These are ignored in the analysis.
Notice: Logging run history failed.
Warning: The following locations have a population totaling zero for the specified date(s).
Location 360559800001, 2009/7/1
Location 360639401001, 2009/7/1

```

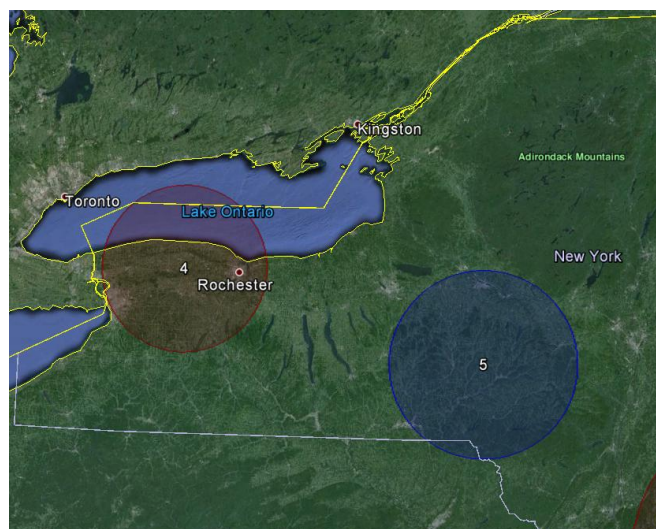
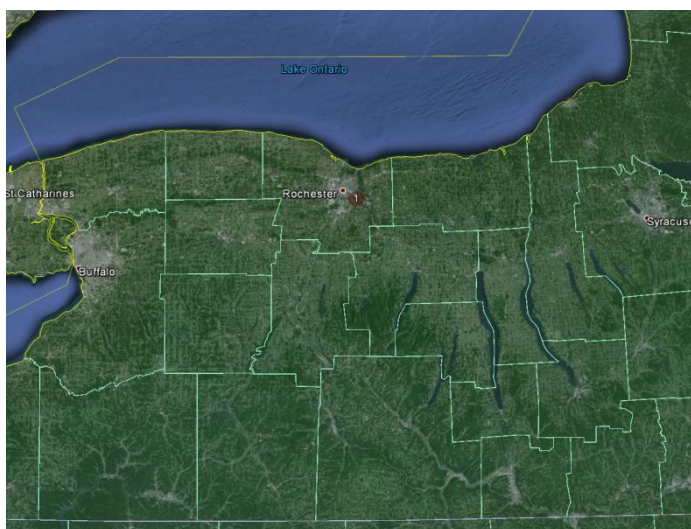
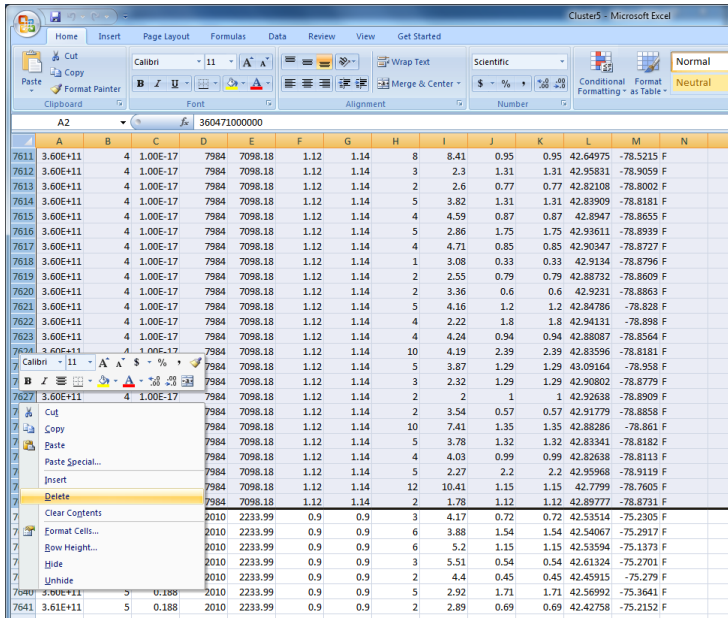


Figure 7: Maps of the sub analysis of Cluster 4 (left) versus Tutorial #1 Analysis including all of New York focused on Cluster 4 and 5 (right)

At this point, please save the parameter files for both these clusters as described in section 1.3. These will be used in future chapters.

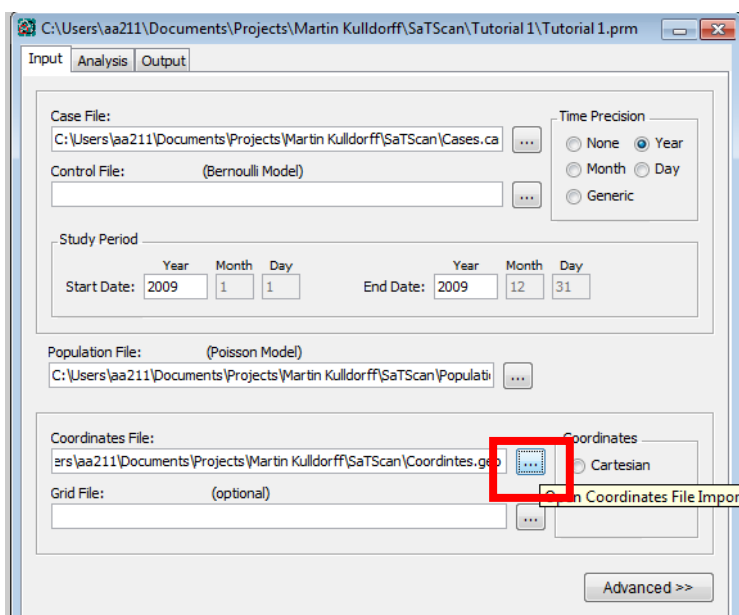
### 3.8. Cluster 5 Sub Analysis

To create the coordinate file for the cluster 5 sub analysis, we will delete all entries that do not have a 5 in column B:



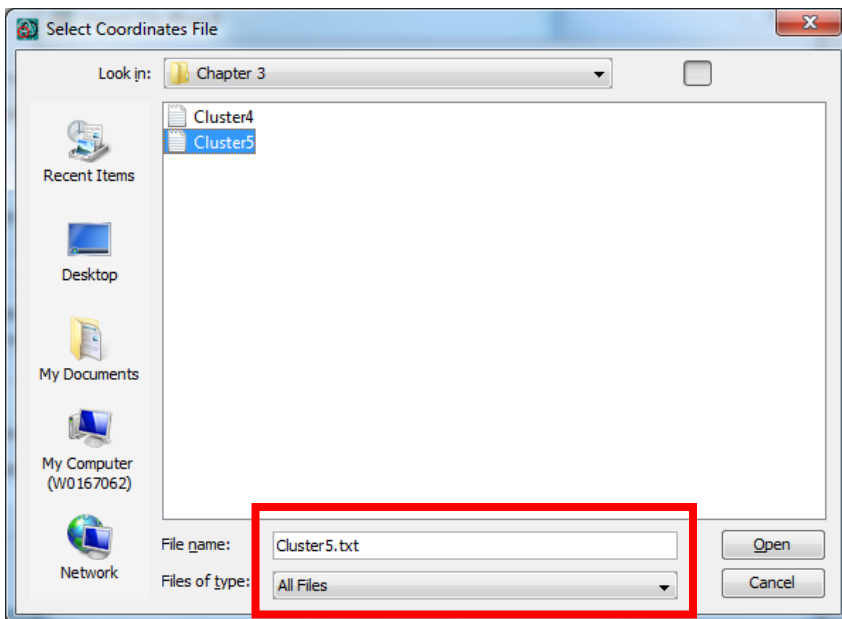
Once this is done, save this excel as a tab delimited text file called “Cluster 5.txt”. Instructions on how to save an excel file as a text file are in section 3.3.

Please open the saved parameter file for tutorial 1, as described in Section 1.5. These new files will be uploaded into SaTScan as the coordinate file through the Import Wizard.

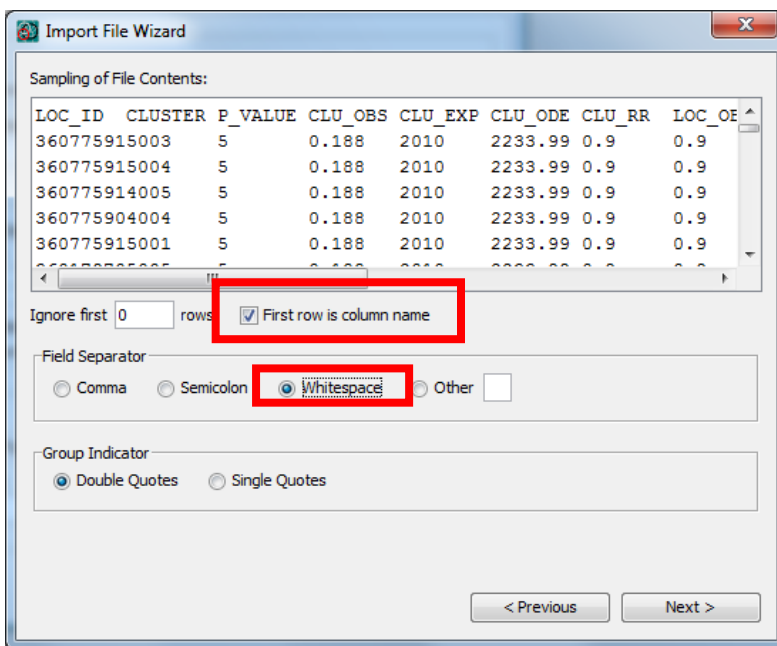




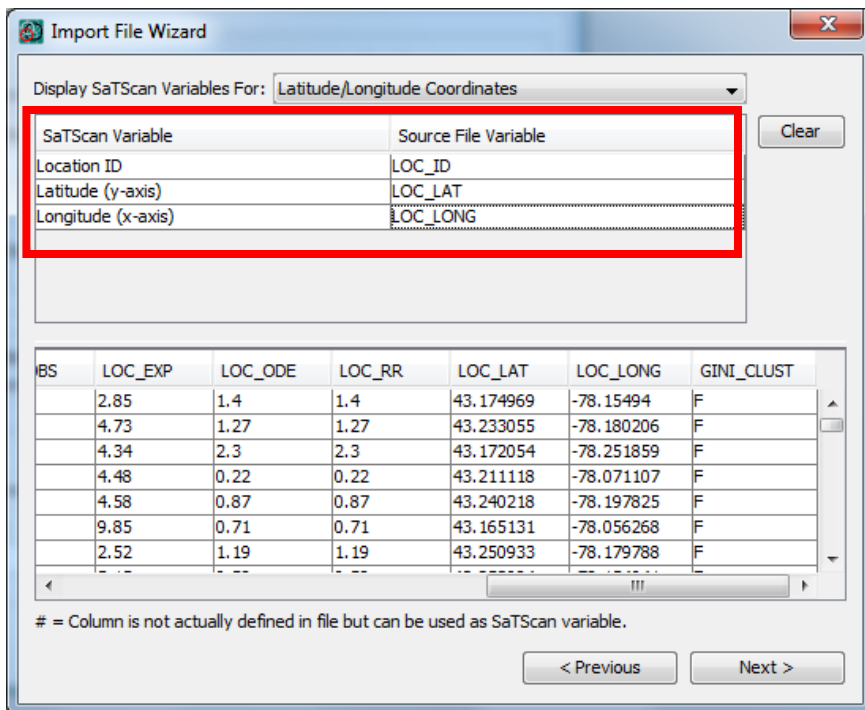
After opening the Import Wizard, make sure to select “All Files” for the “Files of Type”.



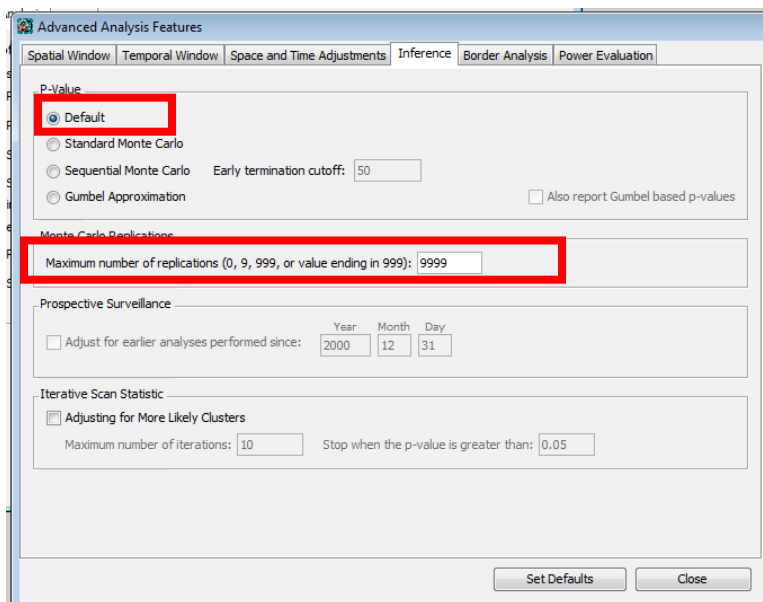
Make sure to select field separator as: ‘Whitespace’ and ‘First row is column name’.



Make sure to select “LOC\_ID”, “LOC\_LAT” and “LOC\_LONG” for Location ID, Latitude, and longitude respectively.



After following the remaining prompts, run the analysis with 9999 Replications in the Default setting for P-Value under the 'Inference' Tab:



Ensure that the geographical data check is off before running. This can be done by starting at the "Input" tab, going to the "Advanced" tab, "Data Checking" then finally clicking "Ignore observations that are outside the specified geographical area".

```

C:\Users\aa211\Documents\Projects\Martin Kulldorff\SaTScan\Tutorial 1\NYS_Cancer\Results.txt

SaTScan v9.4.2

Program run on: Thu May 19 19:13:04 2016

Purely Spatial analysis
scanning for clusters with high or low rates
using the Discrete Poisson model.

SUMMARY OF DATA
Study period.....: 2009/1/1 to 2009/12/31
Number of locations.....: 478
Total population.....: 2234
Total number of cases.....: 2010
Annual cases / 100000.....: 90039.4

CLUSTERS DETECTED
1. Location IDs included.: 360530307003, 360530307002, 360530305022, 360530305012, 360530308002
Coordinates / Radius...: (42.877888 N, 76.785978 W) / 10.05 km
Population.....: 34
Number of cases.....: 36
Expected cases.....: 30.65
Annual cases / 100000.: 164611.2
Observed / expected...: 1.83
Relative risk.....: 1.85
Log likelihood ratio...: 8.567393
P-value.....: 0.155

Warnings/Errors:
Warning: Some records in population file reference a location ID that was not specified in the coordinates file. These are ignored in the analysis.
Notice: Logging run history failed.

```

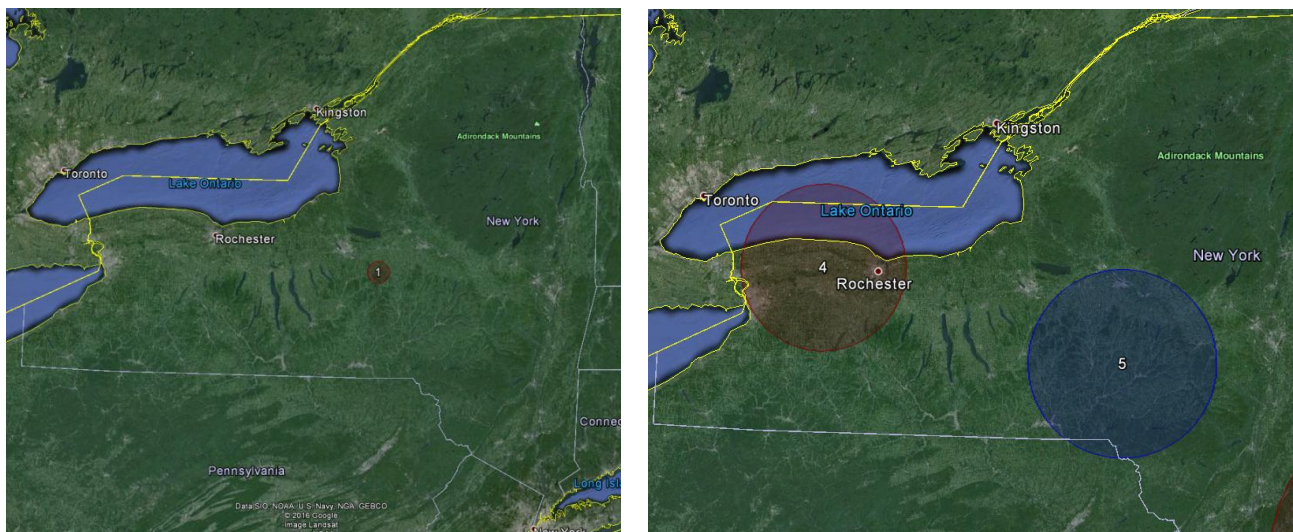


Figure 8: Maps of the sub analysis of Cluster 5 (left) versus the original Tutorial #1 analysis including all of New York focused on Cluster 4 and 5 (right).

### 3.9. Discussion

Within the high cluster 4 sub analysis, there was a significant high cluster of people found with a p-value of 0.017 and a relative risk of 1.48. When compared to the relative risk of 1.11 of the original high cluster 4, we see that there was a higher risk in this part of the cluster. Low cluster 5 sub analysis did not provide any significant results, with the p-value of the most likely cluster being 0.155 with a relative risk of 1.85. Thus, there is a higher incidence of breast cancer within the low cluster, but it is not significant.

# Chapter Four: P-Value and Monte Carlo Replications

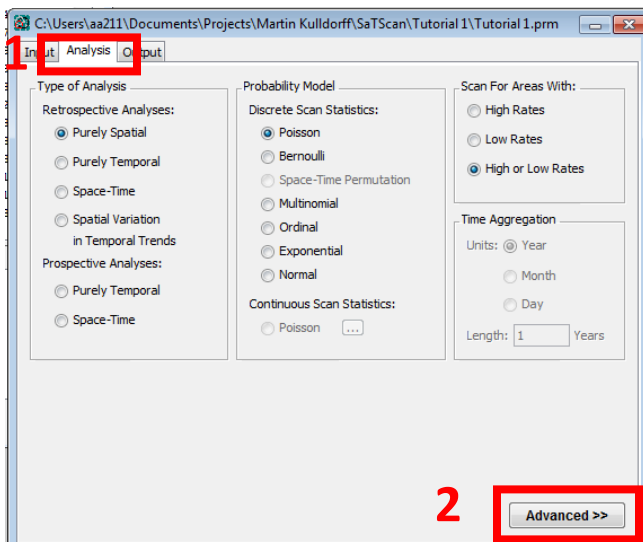
## 4.1. Background Info

For each detected cluster, SaTScan calculates a p-value that is adjusted for the multiple testing of the thousands of circles evaluated. The actual calculation of the p-value can be done in different ways. In this chapter we explore what these options are and how to implement them. To calculate p-values for detected clusters, SaTScan uses computer simulations to create a large number of random replications of the data set generated under the null hypothesis. This is called Monte Carlo hypothesis testing. If the maximum likelihood ratio calculated for the most likely cluster in the real data set is high compared to the maximum likelihood ratios calculated for the most likely clusters in the random data sets, that is evidence against the null hypothesis and for the existence of clusters.

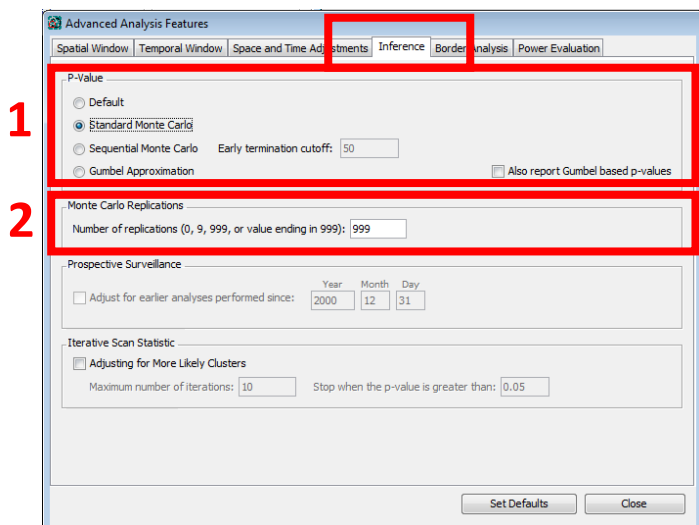
## 4.2. Standard Monte Carlo Hypothesis Testing

The analysis in Tutorial #1 was done using standard Monte Carlo hypothesis testing with 999 Monte Carlo replicates. That is the minimum number of replicates that can be used in SaTScan, but it is sometimes better to use more. We will now show and discuss how and why to do this.

First open the SaTScan session that was saved from Tutorial #1, as described in section 1.4 above. After loading the parameter file from Tutorial #1, switch over to the 'Analysis' tab highlighted below and then the 'Advanced' tab.



After opening the 'Advanced' tab, the 'Inference' tab will show two sections titled 'P-value' and 'Monte Carlo Replications'. In the first section, keep the choice of 'Standard Monte Carlo'. In the second section, change the 'Monte Carlo Replications' from 999 to 9999. After that, click on the green triangle to run the analysis. This analysis will take some time.



Once this has been done for the data for the whole state, please load the parameter file for the low cluster sub analysis from Chapter 3 and redo the same steps.

You probably noted two differences between the 999 and the 9999 analyses. First of all, the latter analysis took much longer to execute. Secondly, the p-values are different. The most likely cluster now has  $p=0.0001$  instead of  $p=0.001$ . This is because a Monte Carlo hypothesis test with  $N$  replications cannot give a p-value smaller than  $1/(N+1)$ . You may also have noticed that other p-values are slightly different. For example, cluster #5 has a p-value of 0.1860 instead of 0.188. The difference in the number of decimals is because of the different number of replications, but the small difference in magnitude is mainly due to chance.

# of Replications	Whole State			Low Cluster Sub analysis from Chapter 3	
	Time (mins)	Cluster #1 (p-value)	Cluster #5 (p-value)	Time (secs)	Cluster #1 (p-value)
999	4.2	0.001	0.188	2	0.143
9,999	17	0.0001	0.1860	4	0.1305
99,999	218	0.00001	0.18519	51	0.12878

Table 3: Summary table for using the Standard Monte Carlo hypothesis testing

Irrespective of the number of Monte Carlo replications, the hypothesis test is unbiased, resulting in a correct significance level that is neither conservative nor liberal nor an estimate. The number of replications does affect the power of the test, with more replications giving slightly higher power. In SaTScan, the number of replications must be at least 999 to ensure excellent power for all types of data sets. In general, the advantage of using more replications is a slight increase in statistical power, at the cost of considerably longer computing time. Normally, it is recommended to use 999 replications for large data sets that take a long time to run, while it is better to use 9999 or 99999

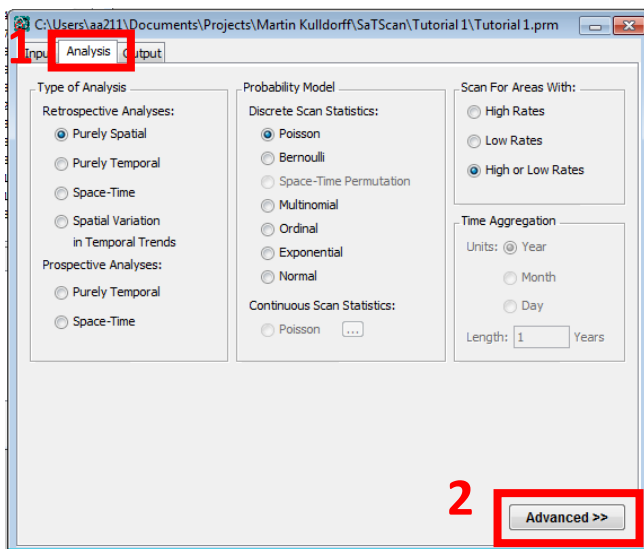
replications for small to medium size data sets that can be run quickly irrespectively of the number of replications.

### 4.3. Sequential Monte Carlo Tests

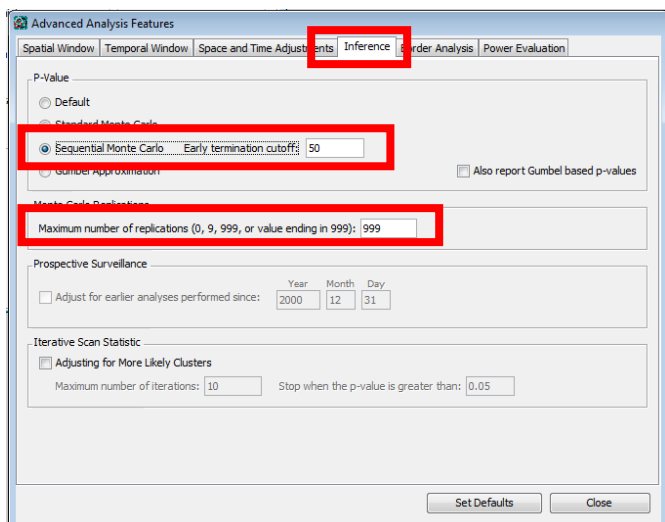
If the p-value is small, then it is important to know exactly how small it is. It is usually inconsequential if the most likely cluster has for example a p-value of 0.352 or 0.376. With the option of sequential hypothesis testing, SaTScan will end the simulations early if it is clear that the most likely cluster will not be statistically significant. We now show how to do that.

For this exercise, we will repeat the sub analysis of cluster 5 from chapter 3, using sequential rather than standard Monte Carlo hypothesis testing. Load the parameter file for the low cluster sub analysis from Chapter 3.

To change to the sequential Monte Carlo based p-value, go back to the 'Advanced' tab under the main 'Analysis' Tab:



Next, choose the 'Inference' tab. There, change the 'P-Value' method to 'Sequential Monte Carlo'. For now, keep the number of Monte Carlo replications to 999. Set the 'Early Termination Cutoff' to 50. This means that the Monte Carlo simulations will stop after there are 50 data sets generated under the null hypothesis with a maximum likelihood higher than the maximum likelihood for the real data set.



Now run the analysis by clicking the green triangular button. You will notice that SaTScan did not do 999 Monte Carlo replications, but stopped after only 323 replications. In the standard analysis in Chapter 4.2, the p-value was 0.143. In the sequential analysis, the p-value is instead 0.155. Both of these are valid p-values, as long as you do not run both analyses and deliberately select the smaller or larger one. In this sense, it is no different from having slightly different p-values when using 999 or 9999 Monte Carlo replications.

Next, change the number of Monte Carlo replications from 999 to 9999, and rerun the analysis. Then do the same with 99999 replications. Did the latter analyses take longer to run? Did you get a different p-value? After how many replications did the simulations stop?

Please import the parameter file for Tutorial 1, as described in Chapter 1 and redo the same steps.

Whole State					Low Cluster Sub analysis from Chapter 3		
# of Replications	Time (mins)	High Cluster 1 (p-value)	High Cluster 5 (p-value)	Terminated after:	Time (secs)	Low Cluster 1 (p-value)	Terminated after:
999	3.2	0.001	0.188	Did not terminate	1	0.155	323 replications
9,999	24	0.0001	0.1860	Did not terminate	1	0.155	323 replications
99,999	152	0.00001	0.18519	Did not terminate	1	0.155	323 replications

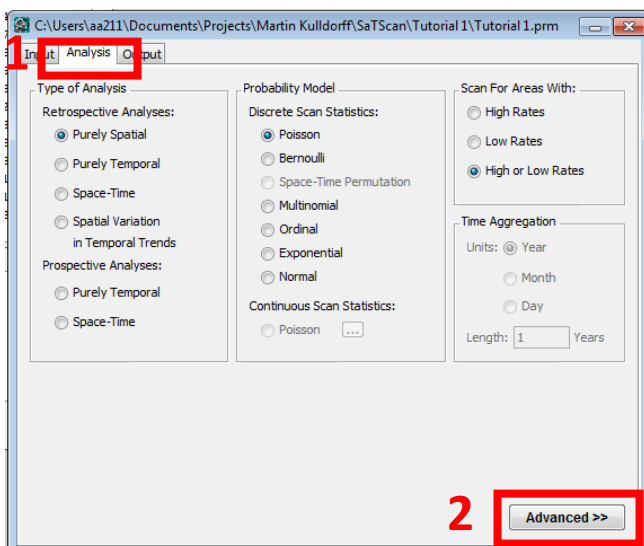
Table 4: Summary table for using the Sequential Monte Carlo inference

#### 4.4. Gumbel based P-values

For very large data sets, it may be too time consuming to run an analysis with more than 999 replications, but it may still be important to have p-values with higher precision than three decimals. In such situation, one can request that SaTScan calculate Gumbel based p-values. While the distribution of the spatial and space-time scan statistics cannot be derived analytically, it is known to follow a Gumbel extreme value distribution (Abrahms et al 2010). To calculate the Gumbel based p-value, SaTScan first runs 999 or whatever number of replications requested by the user. It then fits a Gumbel distribution to those empirical maximum likelihood statistics. This fitted distribution is then used to calculate the Gumbel-based p-value. Note that, unlike the standard and sequential Monte Carlo hypothesis tests, these p-values are not exact, but they are very good approximations.

For this analysis we will go back to using the breast cancer data from the whole state, so reload the parameter file that was saved after Tutorial #1. As with the last example, we will also compare this to the sub analysis of cluster 5 from chapter 3.

To request a Gumbel-based p-value, go back to the 'Advanced' tab under analysis, and then choose the inference tab.

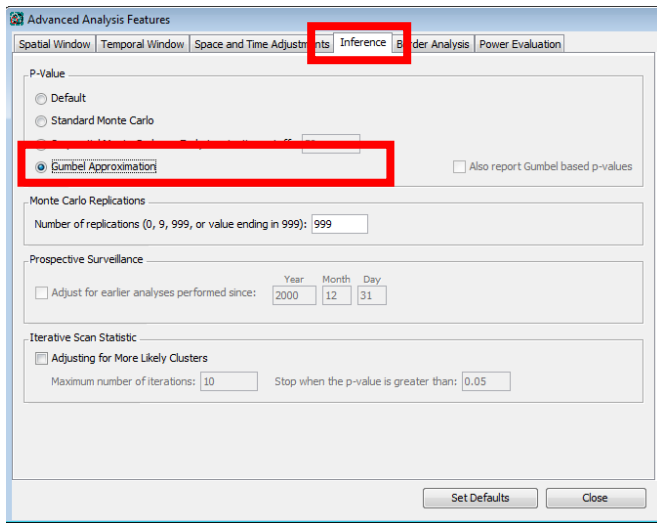


On the inference tab, change the 'P-Value' method to 'Gumbel Approximation'. Then select 999 replications. Then run the analysis by closing the windows and clicking the green play button.

The p-value for the most likely cluster is now  $p < 0.0000001$ , while it was 0.001 with the standard or sequential Monte Carlo hypothesis testing. This was the whole point with the Gumbel approach, to find out if the p-value is actually close to 0.001 or something much smaller.



Note that for cluster #5, Gumbel gave  $p=0.18$  while the standard approach had  $p=0.188$ , reflecting the good approximation of the Gumbel based p-values.



If you want to try the Gumbel approximation for another example, load the parameter file for the low cluster sub analysis from Chapter 3 and redo the same steps.

	Whole state			Low cluster sub analysis from chapter 3	
# of Replications	Time (mins)	High Cluster 1 (p-value)	High Cluster 5 (p-value)	Time (secs)	Low Cluster 1 (p-value)
999	3.3	< 0.0000001	0.18	1	0.14
9,999	17	< 0.0000001	0.18	3	0.13
99,999	157	< 0.0000001	0.18	40	0.13

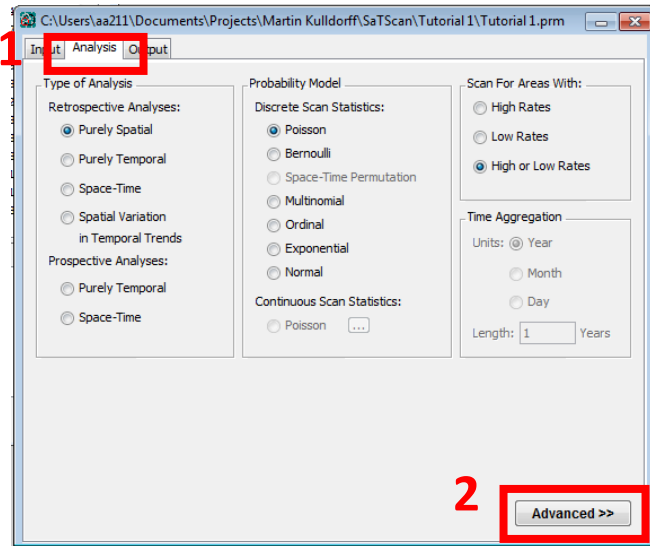
Table 5: Summary table for inferences using the Gumbel approximation.

The Gumbel based p-values are only available for purely spatial and space-time analyses with the discrete Poisson, Bernoulli and space-time permutation probability models. It will soon be available for the ordinal and multinomial models as well. It is not available for other probability models, since it has not yet been evaluated whether the Gumbel approximation works for those analyses.

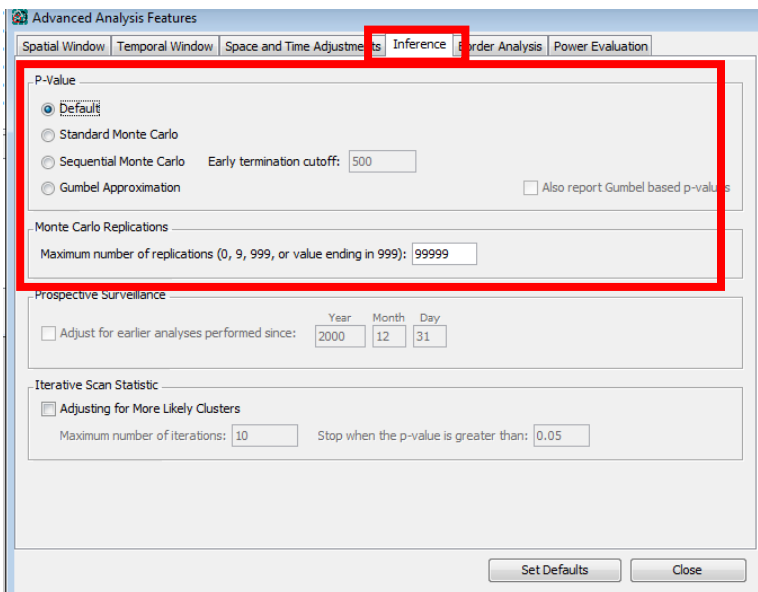
#### 4.5. Default P-value Setting

For the discrete Poisson, Bernoulli and space-time permutation models, the default p-value setting is a combination of the three approaches described above. The sequential version is used if it is able to terminate the analysis early. If the analysis continues to the end, the Gumbel based p-value is used if the p-value is very small while the standard Monte Carlo p-value is presented it provides sufficient precision.

The original dataset from Tutorial #1 was run using the standard Monte Carlo based p-values. First, please make sure to load each parameter file. To change to the default p-value setting, go back to the 'Advanced' tab under analysis, and then select the 'Inference' tab.



From here, change the 'P-Value' method to 'Default Method', while keeping the number of replications at 999. If you want, you can repeat the analysis for a larger number of replications and then run the analysis.



Once you have rerun the analysis been done for the data for the whole state, also rerun the low cluster analysis with the default p-value setting. To do this, load the parameter file for the low cluster sub analysis from Chapter 3 and redo the same steps.

# of Replications	High Cluster Analysis Whole State				Low Cluster Sub analysis		
	Time (mins)	Cluster 1 (p-value)	Cluster 5 (p-value)	Terminated after:	Time (secs)	Cluster 1 (p-value)	Terminated after (replications)
999	4.1	< 0.0000001	0.188	Did not terminate	0	0.155	323
9,999	19	< 0.0000001	0.1860	Did not terminate	2	0.1293	3868
99,999	227	< 0.0000001	0.18519	Did not terminate	18	0.12801	39059

Table 6: Summary table for using Default p-value inference

Looking at Table 6 it is evident that the p-value for cluster #1 in the Whole State Analysis was calculated using the Gumbel approximation, while the p-value for Cluster #5 was calculated using the standard Monte Carlo approach. As a contrast, in the low cluster sub analysis, all p-values were calculated using sequential Monte Carlo hypothesis testing. This can be verified by comparing the values found in this table to the prior tables in this chapter.

## Chapter Five: Maximum Cluster Size to Evaluate

### 5.1. Background

SaTScan will search for clusters at many different locations and for many different geographic sizes. A cluster is never allowed to contain more than 50 percent of the population at risk though. The reason for this is that a larger size, such as e.g. 90%, is more appropriately interpreted as a lower disease rate in the 10% of the area outside the 'cluster' rather than as an excess disease rate covering almost the whole study region. While 50 percent is the default maximum, a smaller maximum may be requested. As a set of advanced features, a more restrictive upper limit on the cluster size can be specified in one of the following three ways: (i) as a percentage of the population at risk that is smaller than 50 percent, (ii) as a radius of the circular cluster expressed in kilometers, or (iii) as a percentage of a special 'population' specified a separate input file, where the 'population' can be very different from the underlying population used to calculate expected counts.

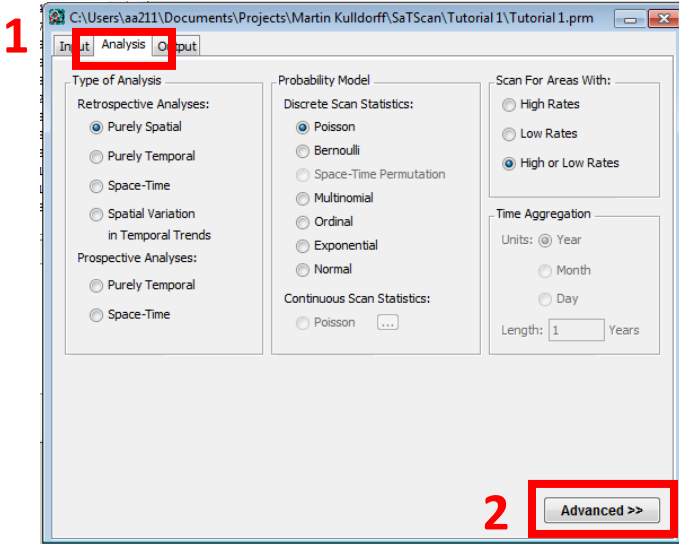
In general, one should pick the maximum geographical cluster size in such a way that any clusters that are bigger have no clinical, public health or scientific interest. For example, if one is looking for clusters of pertussis in the United States, in order to find local outbreaks, a cluster that covers 40 percent of the United States population is too big to have a meaningful interpretation as a localized disease outbreak.

A key feature of the spatial scan statistic and the SaTScan software is that it adjusts the analysis for the multiple testing inherent in the many different potential cluster sizes evaluated. For this to work, one should never run multiple analyses with different maxima. If that is done, it is only the analysis with the larger maximum that is valid, as it is the only analysis that adjusts for both the smaller and larger clusters sizes that was actually looked at. What may sometimes be interesting is to find both smaller and larger clusters, but this is accomplished by selecting the type of clusters that are reported in the results, to be covered in the next two chapters.

### 5.2 Maximum as a Percentage of the Population at Risk

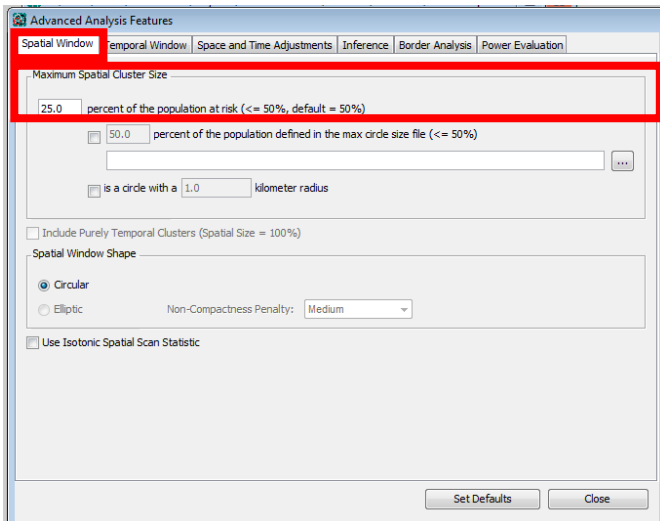
In SaTScan Tutorial #1, the maximum cluster size was set to 25 percent of the underlying population at risk. That is, clusters were only evaluated if the expected number of breast cancer cases were less than 25 percent of the total number of breast cancer cases in the state. Potentially, it could be argued that only smaller clusters are of public health importance, and it would then be reasonable to choose a smaller maximum of say 10 percent of the population at risk. It will now be shown how to do that.

First open the SaTScan session that was saved from Tutorial #1, as described in section 1.4 above. After loading the parameter file from Tutorial #1, switch over to the 'Analysis' tab highlighted below and then click on the 'Advanced' button located at the bottom right.



The screen below should appear. One set of advanced features are related to the spatial aspects of the scanning window, and these are located on the 'Spatial Window' tab. One of the features on this tab is the 'Maximum Spatial Cluster Size' section highlighted below.

In the top left corner of the 'Maximum Spatial Cluster Size', change the percent of the population at risk from 25% to 10%. After that, close the tab, and run the analysis.



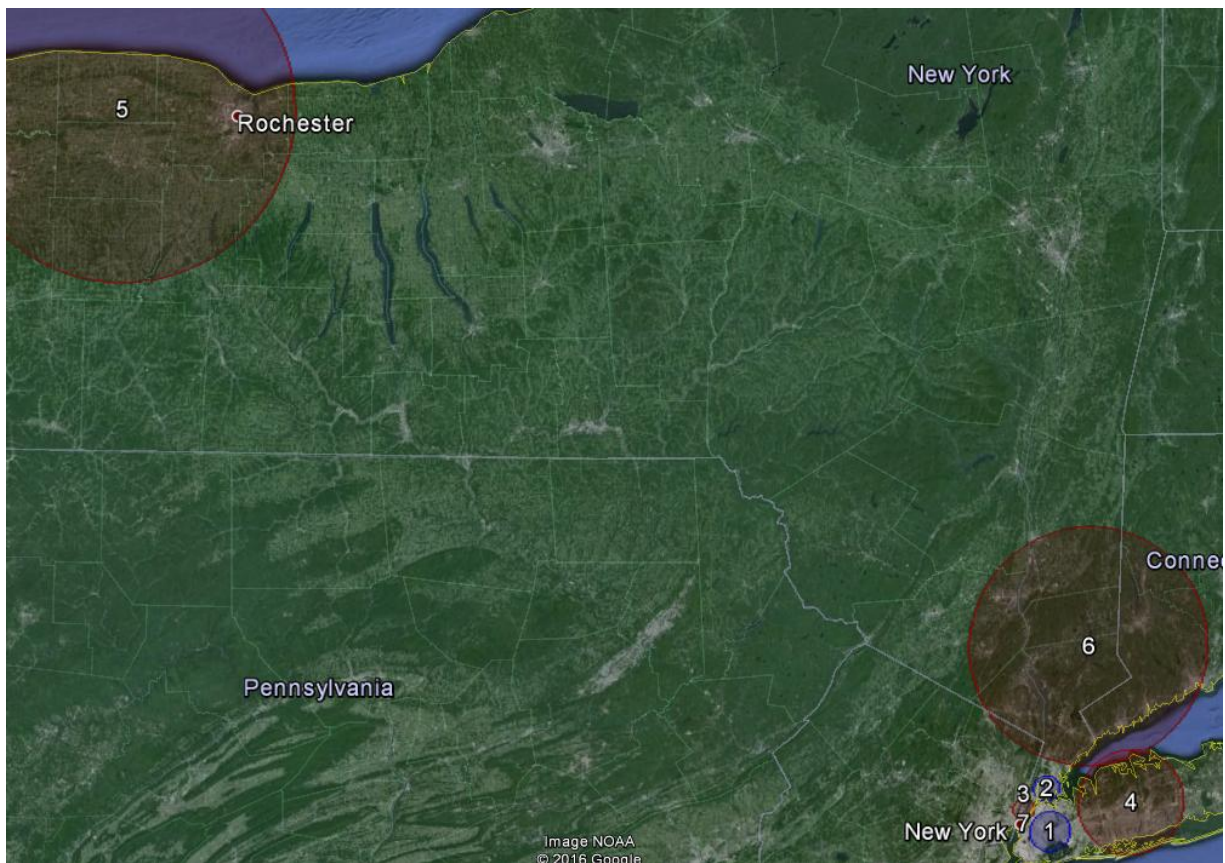
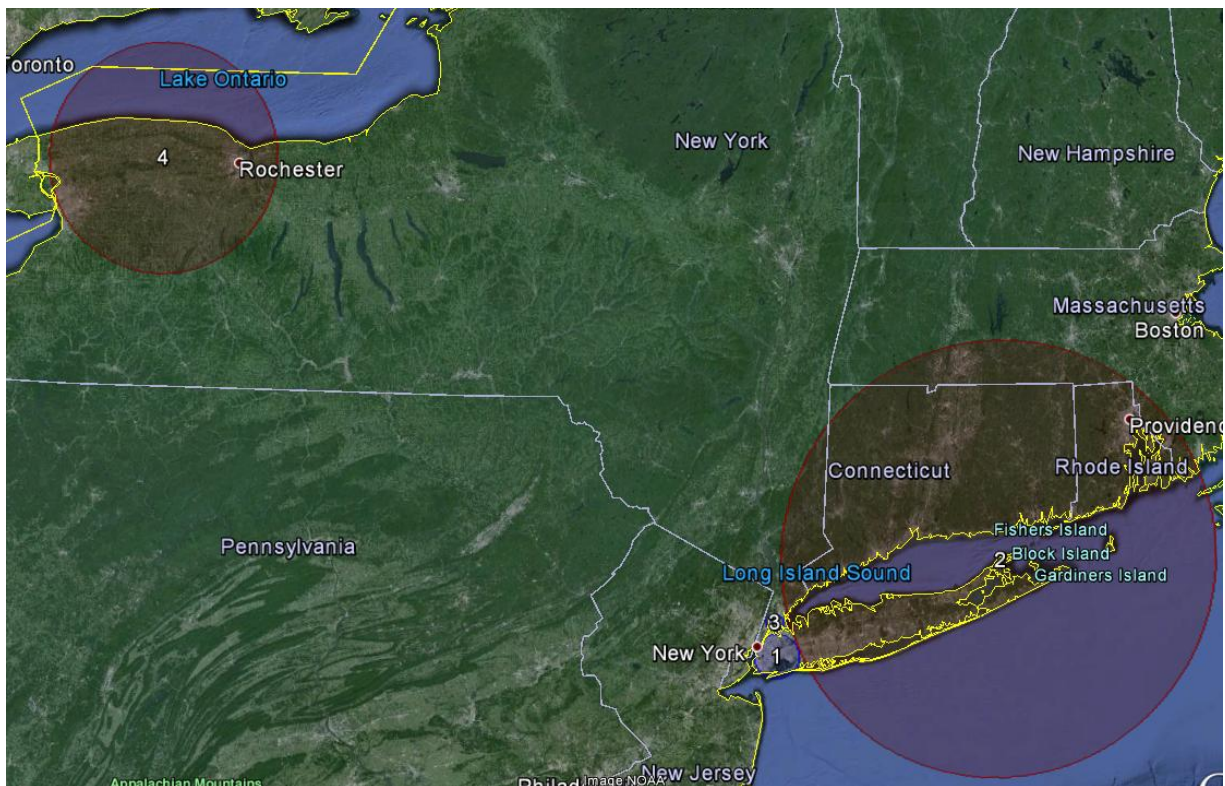


Figure 9: Breast cancer incidence clusters in New York State when the circular spatial scan statistic was run with a 25% (top) versus a 10% (bottom) maximum cluster size.

25%						10%					
Cluster	Radius (km)	Observed Cases	Expected Cases	Relative Risk	p-value	Cluster	Radius (km)	Observed Cases	Expected Cases	Relative Risk	p-value
1	12.82	13642	15886	0.83	< 0.0000001	1	7.80	5901	7229	0.80	< 0.0000001
2	125.47	13416	15019	1.15	< 0.0000001	3	4.08	3648	2974	1.24	< 0.0000001
						4	20.41	7831	6869	1.16	< 0.0000001
						6	45.93	6330	5684	1.12	0.0000000000016
3	4.97	3112	3976	0.77	< 0.0000001	2	4.97	3112	3976	0.77	< 0.0000001
4	65.97	7984	7098	1.14	< 0.0000001	5	65.97	7984	7098	1.14	< 0.0000001
5	74.03	2010	2234	0.90	0.188	8	74.03	2010	2234	0.90	0.185

Table 7: Breast cancer incidence clusters in New York State when the circular spatial scan statistic was run with a 25% versus a 10% maximum cluster size.

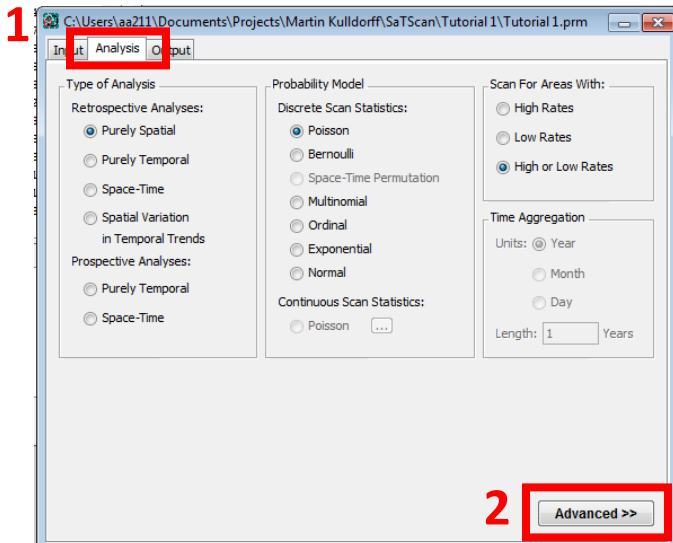
Figure 9 and Table 7 show the results of the new analysis with the 10% maximum, as well as the results of the prior Tutorial #1 analysis with a 25% maximum for comparison. Some of the clusters are identical, such as the one in the northwestern part of the state around Buffalo and Rochester (#4/5). Other clusters have changed. With a 25% maximum, there was a large cluster (#2) covering central and eastern Long Island as well as the southeastern part of the Hudson River Valley. With a 10% maximum, this cluster was split into two, one for central Long island (#4) and another for southeastern Hudson River Valley (#6). Another difference is the size of the cluster in Brooklyn and Queens (#1). With a 25 percent maximum, the detected cluster has 15886 expected cases, which is 22% percent of the total. This cluster is too large when the maximum is set to 10 percent of the population at risk, forcing the cluster to be smaller. Another important feature to note is the difference between the p-values recorded for cluster 5 with a 25% maximum and cluster 8 (not shown in map, because it is not statistically significant) from the analysis with a 10% maximum. As expected, the p-value for these identical clusters is smaller and more precise (0.185 as opposed to 0.188) for the 10% maximum because there is less multiple-testing done.

Since the set of detected clusters are different when different maxima are used, which clusters are the correct ones? The answer is both sets and neither. When using the spatial scan statistics, the exact borders of the detected clusters are uncertain. For any detected cluster, there are probably some areas within the cluster that do not have an excess risk of the disease and there are most likely some areas outside the cluster that do have an excess risk. The spatial scan statistic only provides the rough location and size of the clusters. When comparing the two analyses above, the key thing is that both results are rather similar in terms of which areas have a higher risk and which areas have a lower risk of breast cancer.

### 5.3 Maximum Cluster Size by Geographical Size

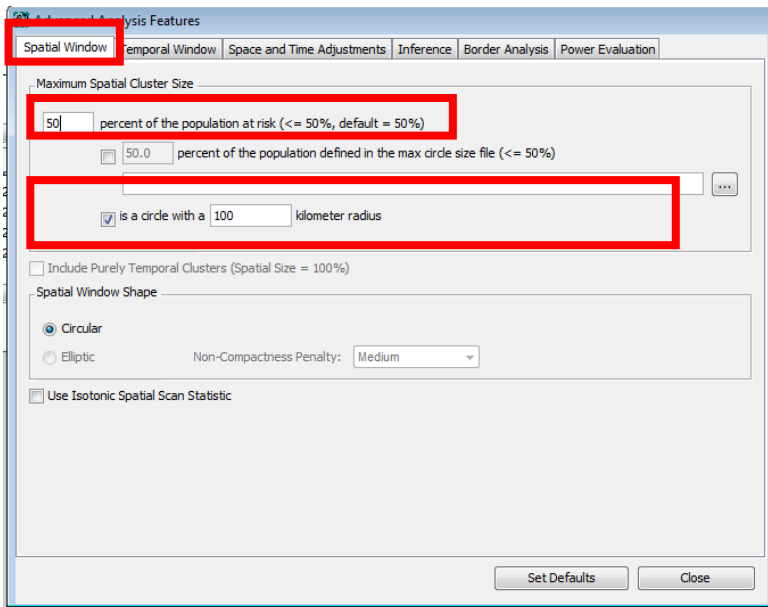
In SaTScan, the maximum cluster size cannot only be specified in terms of the population at risk, but also in terms of its geographical size. The latter is done by specifying the maximum radius of the cluster circle. If latitude/longitude coordinates are used, then the maximum radius should be specified in kilometers. If the standard Cartesian coordinates that are taught in High School are used, the maximum radius should be specified in the same units as the Cartesian coordinates.

For the breast cancer incidence data, suppose we want the maximum cluster size to be a circle with a radius of 100 kilometers. After loading the parameter file from Tutorial #1, switch over to the 'Analysis' tab highlighted below and then click on the 'Advanced' button located at the bottom right.

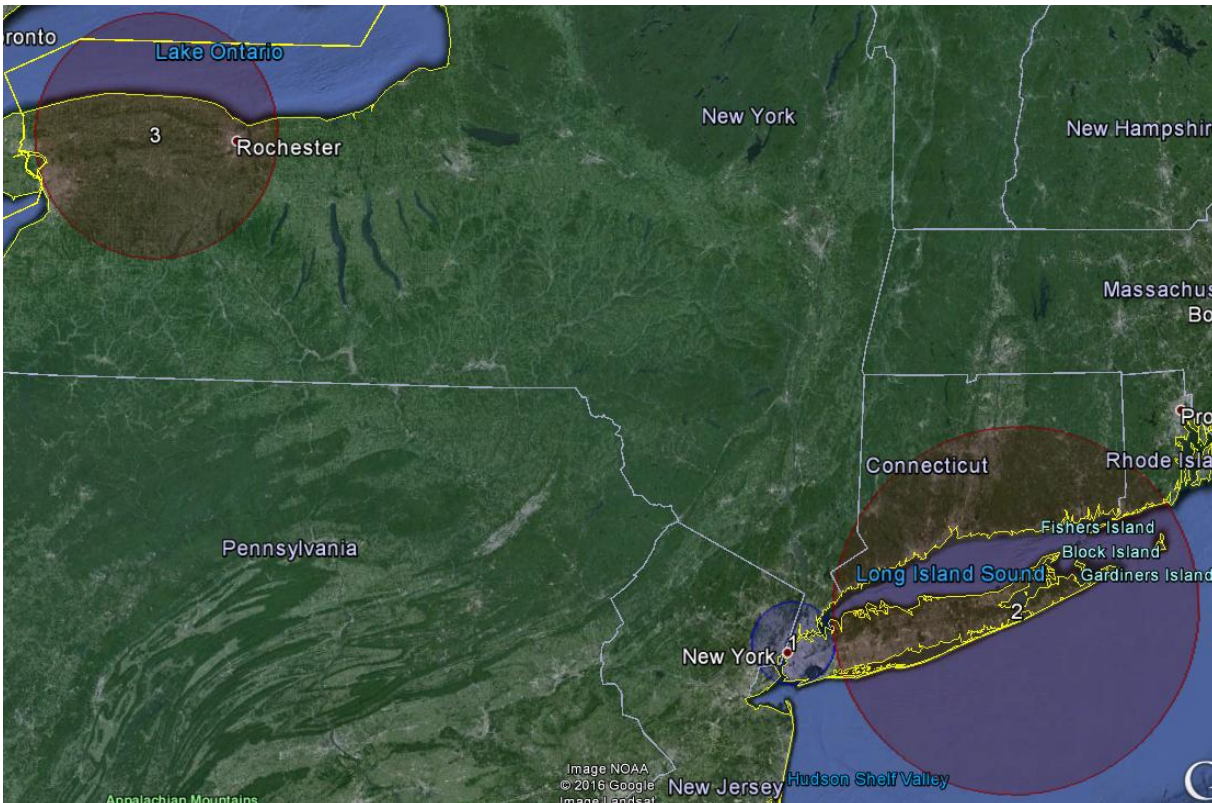


The screen below should then be shown. The 'Spatial Window' tab relates to the spatial aspects of the scanning window. One of the features on this tab is the 'Maximum Spatial Cluster Size' section highlighted below.





In the top left corner of the 'Maximum Spatial Cluster Size'. First check the box at the bottom of this section, and then write 100 as the maximum "kilometer radius". To ensure that the 100km radius is not restricted, change the percent of the population at risk from 25% to 50%, the largest possible cluster size. This ensures that the output truly represents the parameter we have specified in this tutorial. After that, close the tab, and run the analysis. The results of this can be seen below, compared to 25% of population at risk:



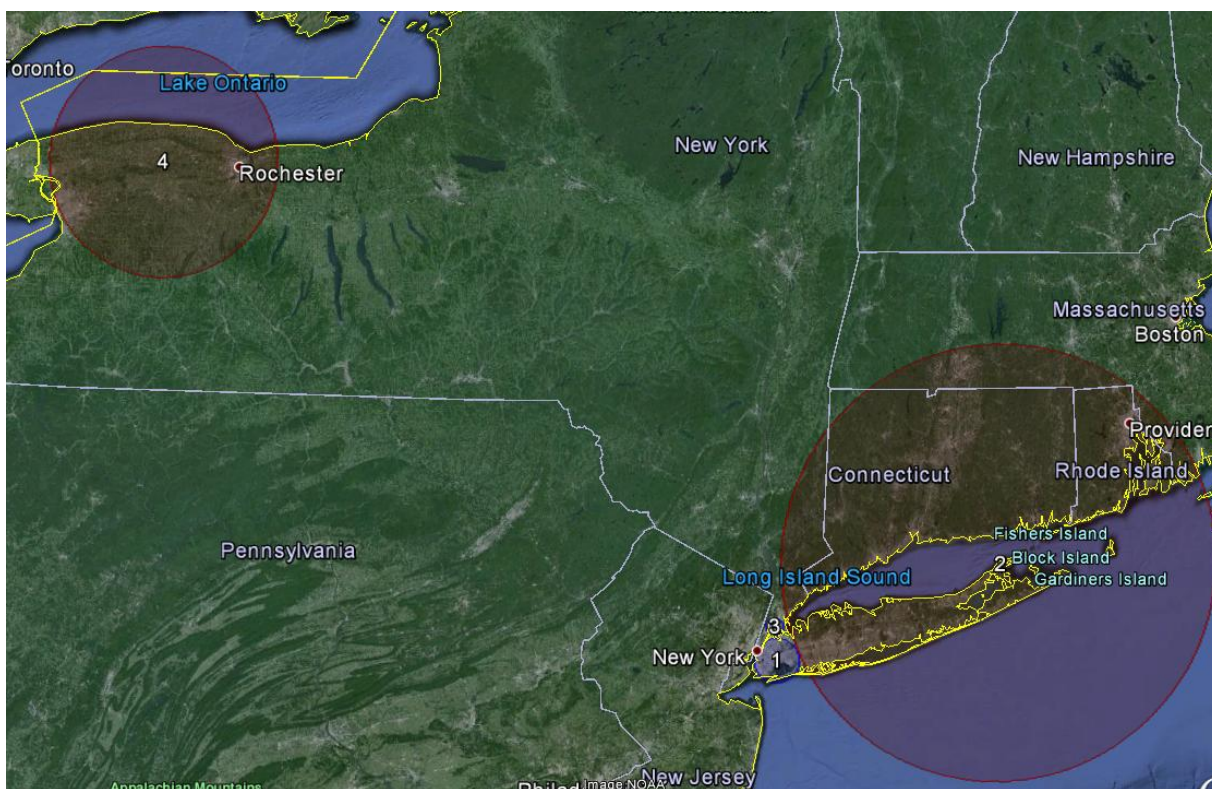


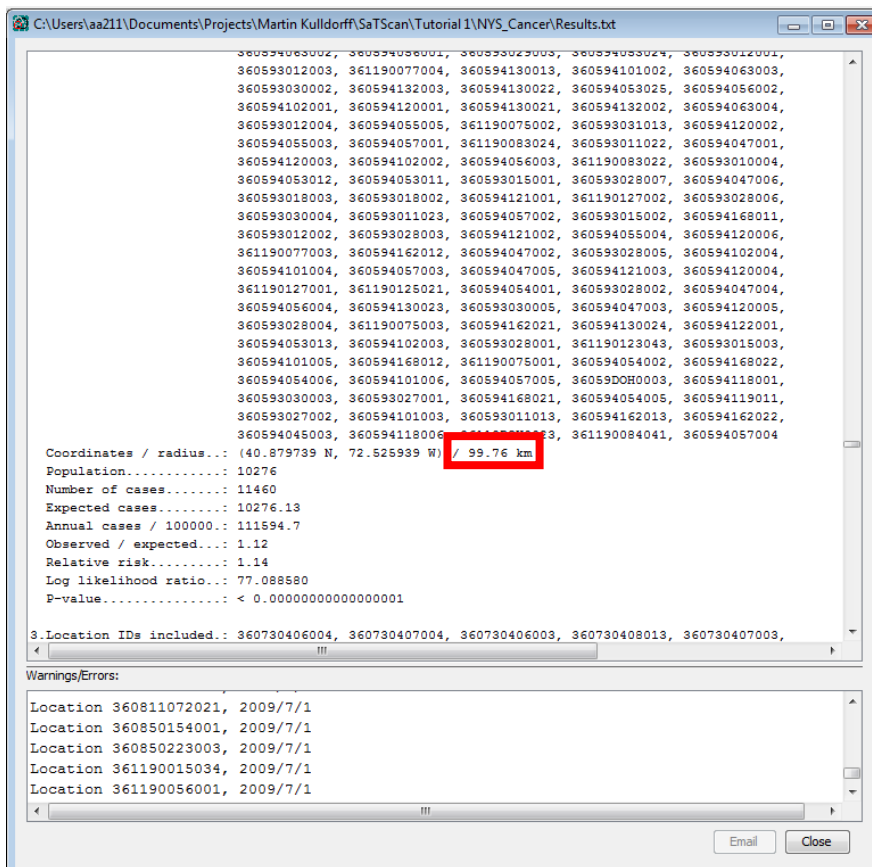
Figure 10: Breast cancer incidence clusters in New York State when the circular spatial scan statistic was run with a 100% (top) versus a 25% maximum cluster radius size (bottom).

25% Maximum				100 km Maximum			
Cluster #	Maximum Radius (km)	Observed Counts	Expected Counts	Cluster #	Maximum Radius (km)	Observed Counts	Expected Counts
1	12.82	13642	15886	1	22.41	25954	28716
3	4.97	3112	3976				
2	125.47	15019	13416	2	99.76	11460	10276
4	65.97	7984	7098	3	65.97	7984	7098

Table 8: Breast cancer incidence clusters in New York State when the circular spatial scan statistic was run with a maximum cluster size of 100km versus 25%.

Of the significant clusters from the original Tutorial #1 analysis presented in the above images, cluster 2 has a radius of 125.47km, which is too big when the maximum is set to 100km. Instead, the new analysis finds a slightly smaller cluster with a 99.76km radius.

This size can be seen in the output shown here:



The top cluster in the new analysis has a radius of 22.41km. The expected number of cases in this cluster is 28716. Since this is a larger than 25%, that cluster could not be found with the 25% population in the analysis maximum, which instead found two smaller clusters.

#### 5.4 Maximum Cluster Size using a Special Max Circle Size File

Suppose that we want to maximum cluster size to correspond to 10% of the population in New York State. That is slightly different from what was done in Chapter 5.2, since the maximum in that case was 10% of the population at risk. The latter only counts women, since the analysis is done for female breast cancer. Moreover, the population at risk reflects the expected counts of breast cancer, which is different from the actual number of women, since older women are at higher risk for breast cancer compared to younger women.

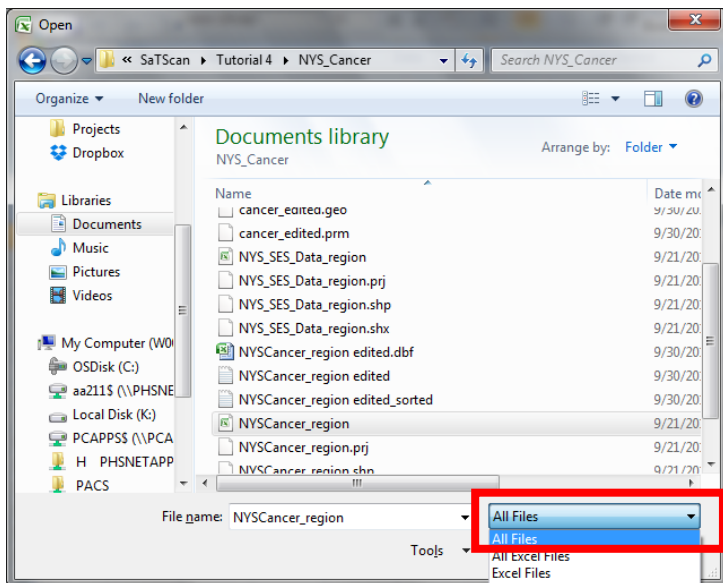
SaTScan provides the option to define the maximum circle size using a different population that is used to calculate the expected counts than the population at risk. This is done by specifying a different population in a special max circle size input file and by requesting that this file be used to define the maximum cluster sizes. If this file contains the total population in each area, then the maximum cluster size would be defined in terms of the total population. It can also be used to define the maximum for any other type

of *'population'*. For example, if it contained the number of cats in each area, it could be used to define the maximum in terms of the clusters cat population, although it is hard to imagine why one would want to do that in a study of breast cancer incidence.

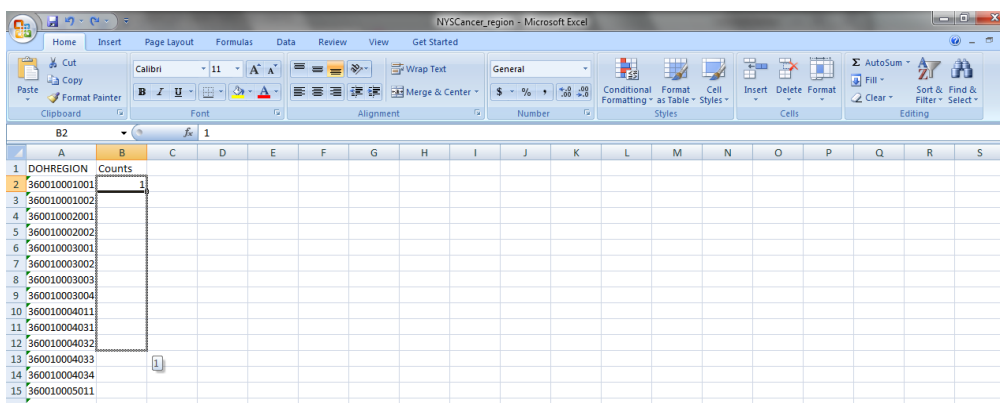
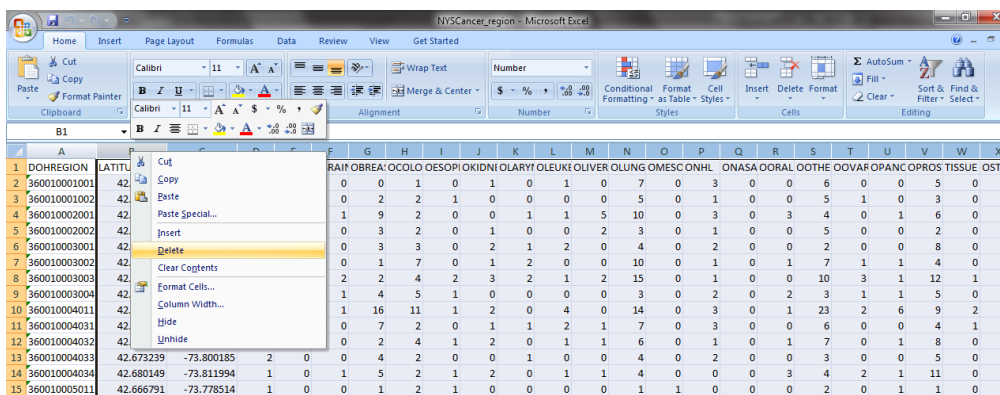
In some situation, one may want to define the maximum circle size in terms of a maximum number of counties, census tracts or some other administrative unit. The special max circle size file can be used for that as well, by setting the *'population'* of each unit to be equal to one. For our breast cancer incidence data, the areas used are Department of Health Regions (DOH Region). We will now show how to set the maximum cluster size to 10% of the DOH Regions in the dataset, which corresponds to 1384.8 DOH regions since there are 13848 total DOH regions.

To use the *'Max Circle Size File'* Option, a new file must be created. This is done by selecting all the location IDs from the geographical location file and creating another column with a corresponding count with a value of 1 for each DOH region.

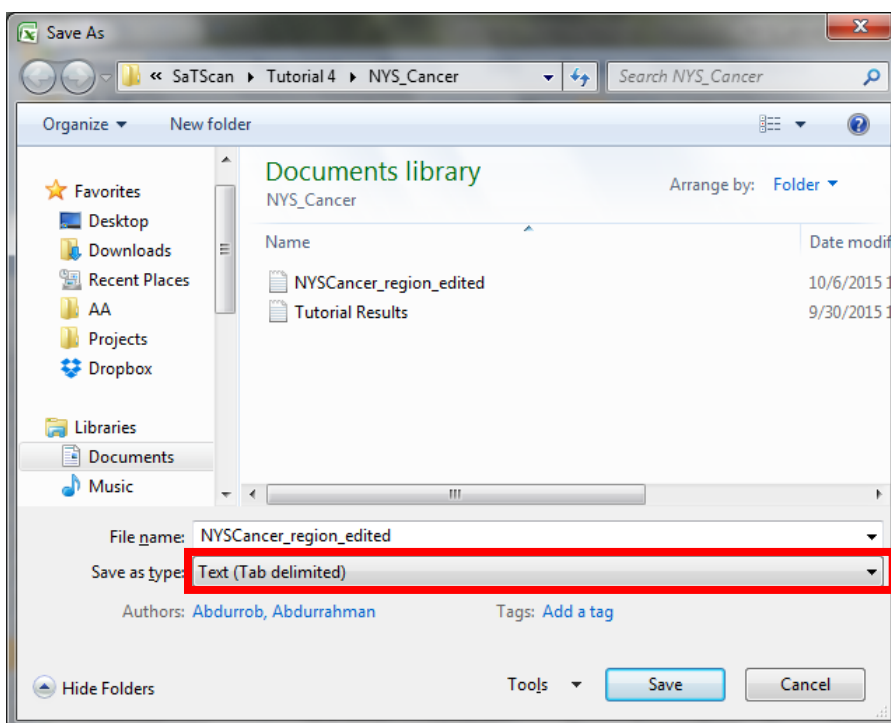
Open Microsoft Excel or a similar program and then open the file NYSCancer\_region.dbf. If you use Microsoft Excel, make sure to change the selection from Excel Files to All Files, so that you will be able to see the dbf file.



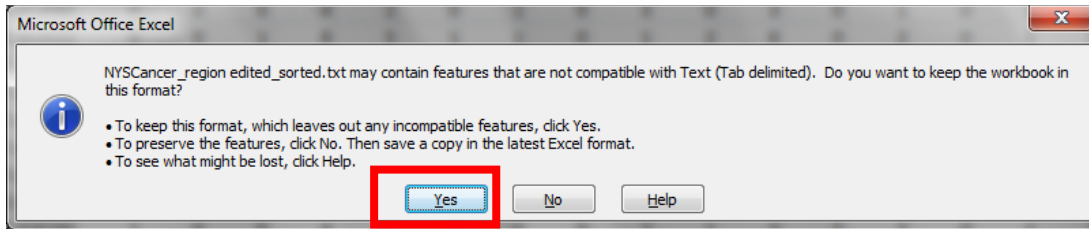
First erase all columns except column A. After this, create a second column called the *'Counts'* column. In the second row, put the value 1. Then set the value to *'1'* for every corresponding DOH Region by dragging the first entry all the way to the end, so that the second column has a 1 in each row.



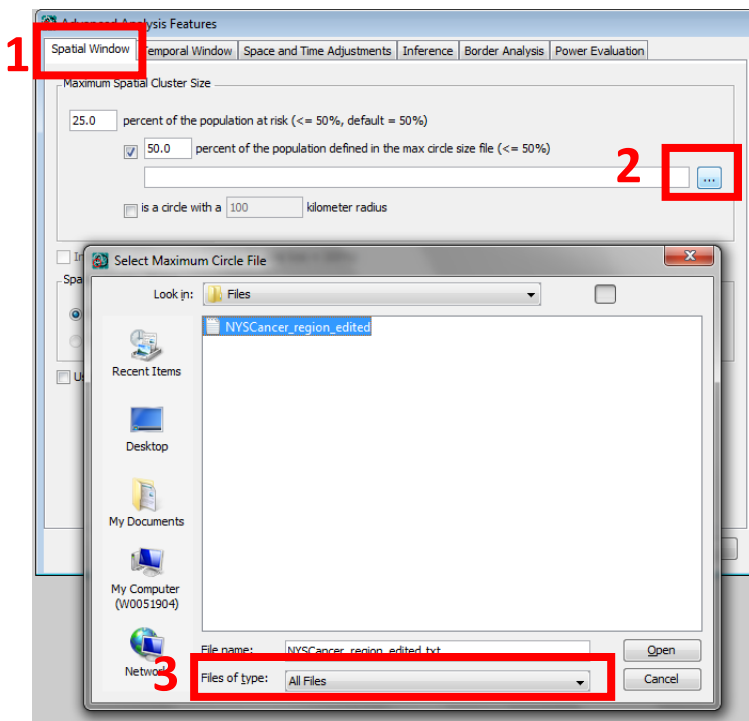
Once the database file has been successfully edited, save the file as a new file entitled: 'NYS\_Cancer\_region\_edited'. Save the file as a tab delimited text file. This will allow SaTScan to read the data file as distinct columns and rows.



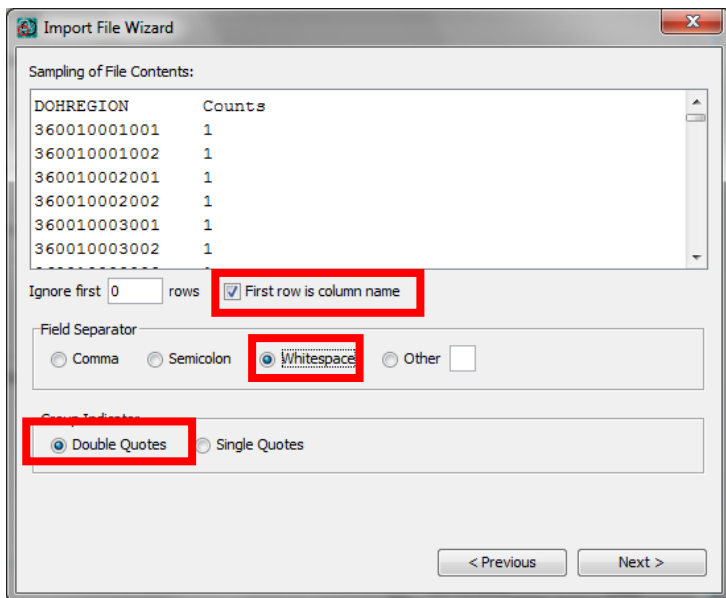
Make sure to select yes, to maintain compatibility:



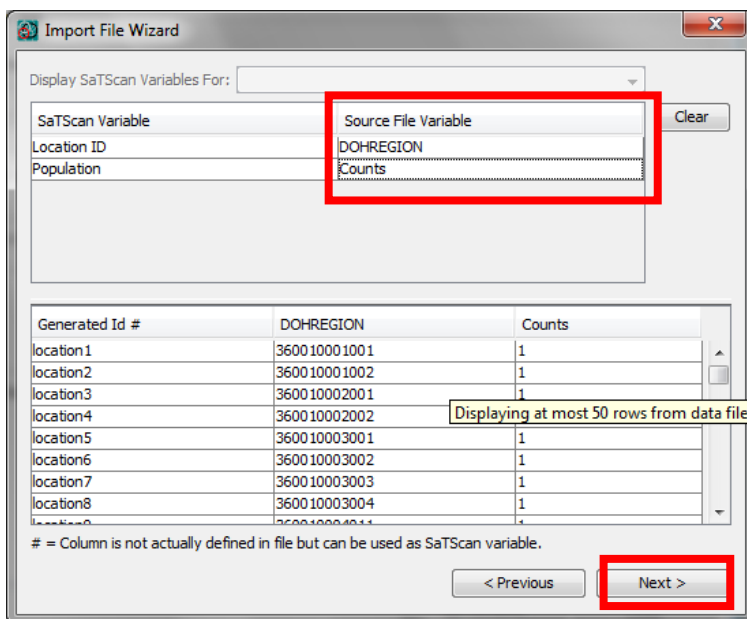
After creating the max circle size file, go back to the SaTScan software and the 'Spatial' Windows tab on the advanced analysis options. Put a check mark on the second row in the 'Maximum Spatial Cluster Size' box, indicating that you want to use a special max circle size file. As the next step click on the box at the very right to upload the newly created file. When uploading it is important to change the 'Files of Type' to 'All Files', otherwise the Maximum Circle File will not be visible.



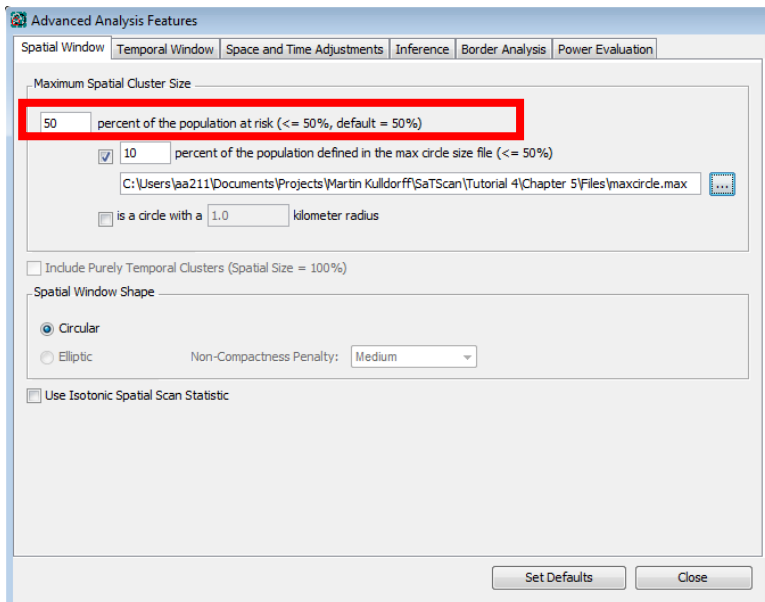
Selecting 'First row is column name', 'whitespace' and 'double quotes' allows the file to be imported correctly with the Import File Wizard:



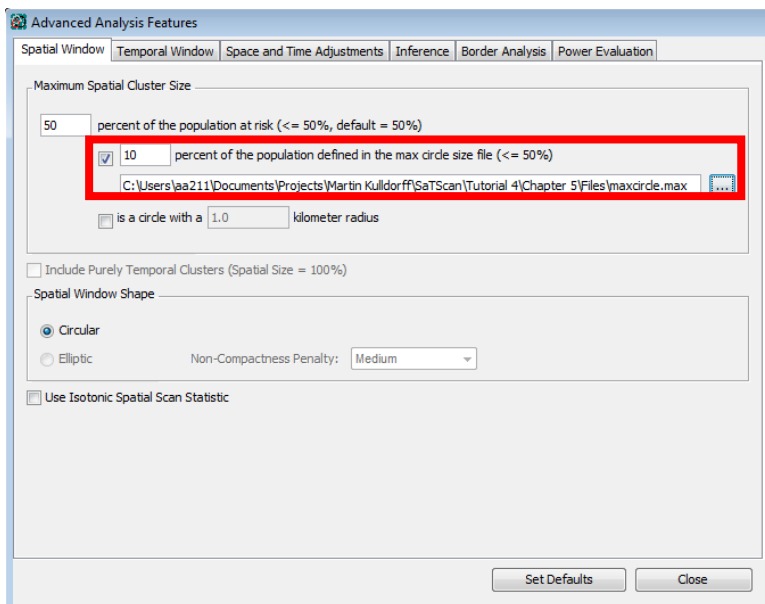
Make sure to specify DOHREGION as the Location ID, Counts as the Population. Then proceed with next and follow the prompts until the Import File Wizard closes (below) and returns you to the *'Spatial Window'* tab.



Since we no longer want a restriction of 25% of the maximum cluster size in terms of the population at risk, change that number to 50%, which represents the largest possible value.



It is now time to define the maximum in terms of the percentage of the DOH regions. Enter this number on the second row in the 'Maximum Spatial Cluster Size' box.





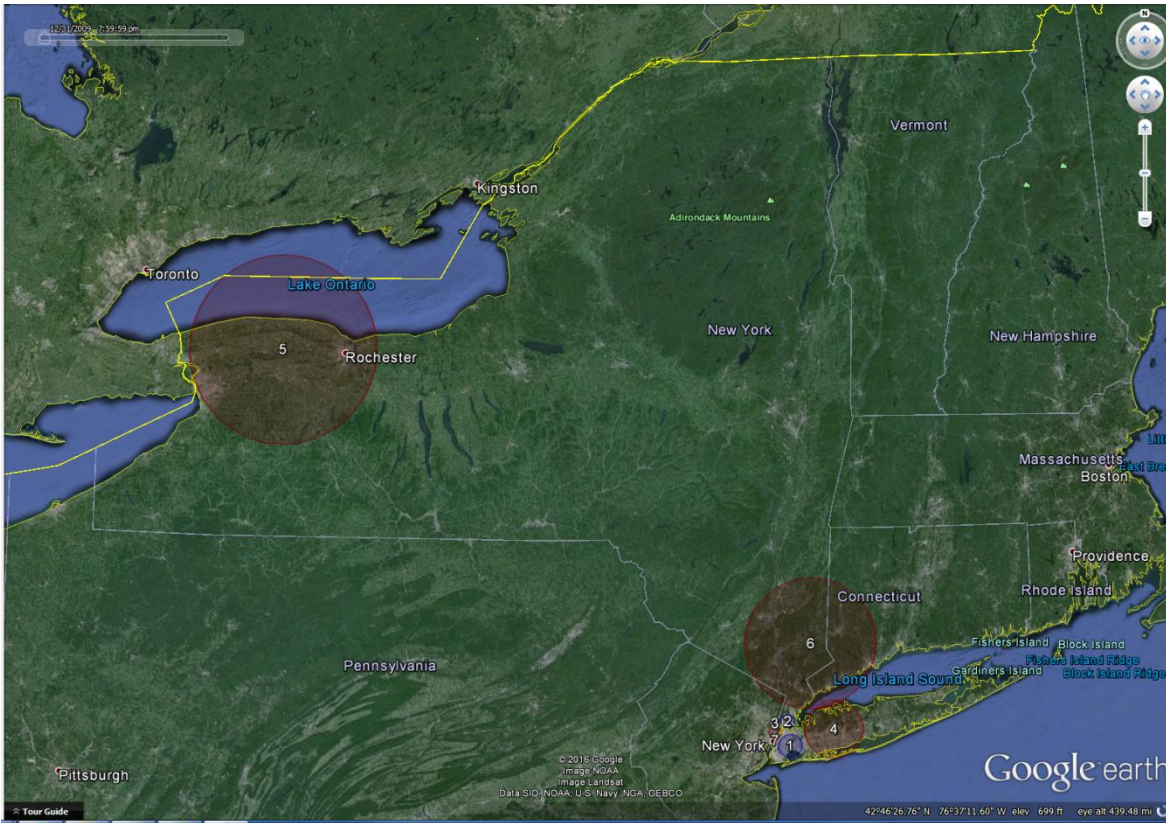


Figure 11: Cluster of breast cancer incidence when the cluster size was set to include a maximum of 10% of the Department of Health Regions.

Figure 11 show the results of the new analysis with a 10% maximum for the number of DOH regions to be included in a cluster. As noted before, this corresponds to a maximum of 1384 DOH Regions. A summary of the clusters in this analysis have the following number of DOH Regions per cluster:

Maximum cluster size of 10% of DOH regions						Maximum cluster size of 10% of population at risk					
Cluster	# of DOH Regions	Observed Cases	Expected Cases	Relative Risk	p-value	Cluster	# of DOH Regions	Observed Cases	Expected Cases	Relative Risk	p-value
1	1384	5953	7289	0.80	< 0.0000001	1	1379	5901	7229	0.80	< 0.0000001
2	775	3112	3976	0.77	< 0.0000001	2	775	3112	3976	0.77	< 0.0000001
3	531	3648	2974	1.24	< 0.0000001	3	531	3648	2974	1.24	< 0.0000001
4	1332	7831	6869	1.16	< 0.0000001	4	1332	7831	6869	1.16	< 0.0000001
5	1349	7984	7098	1.14	< 0.0000001	5	1349	7984	7098	1.14	< 0.0000001
6	1023	6330	5684	1.12	0.00000000000016	6	1023	6330	5684	1.12	0.00000000000016
7	38	265	167	0.63	0.000031	7	38	167	265	0.63	0.000031

Table 9: Comparison of 10% of population at risk versus 10% of the DOH regions

When comparing to Figure 11 to Table 9 with a 10% maximum on the population, it is clear that the results are very similar. For this example restricting the population corresponds to the restricting the DOH region fairly well, but this will not always be the case. Nonetheless, we do see that cluster 1 differs between the two. Both these clusters reach their respective upper boundry: 10% of the total population of 72296 = 7229 people and 10% of 13848 DOH regions = 1384. While the observed cases are slightly different,

there are no tangible differences in relative risks and the p-values of 0.80 and  $<0.0000001$ , respectively.

## Chapter Six: Spatial Clusters to Report

### 6.1 Background

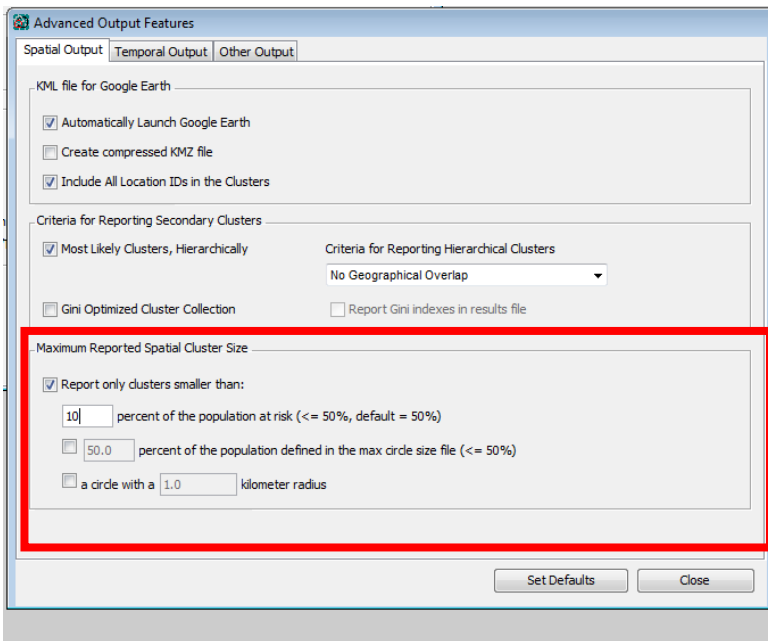
The spatial scan statistic evaluates thousands or millions of different potential cluster locations and sizes, all of which overlap with other potential clusters. In addition to the most likely cluster, there are almost always secondary clusters with almost as high likelihood that greatly overlap with the most likely cluster. This is because removing or adding a small area at the border of the cluster will not greatly change the likelihood of the cluster. SaTScan does not report all of these clusters, since many of them are almost identical, but their existence is a reflection of the fact that the cluster boundaries are uncertain. In SaTScan, there are various advanced options available to explore this uncertainty. In this chapter we cover a few of those, looking at different ways in which the user can specify which of the evaluated clusters to report.

### 6.2 Maximum Reported Spatial Cluster Size

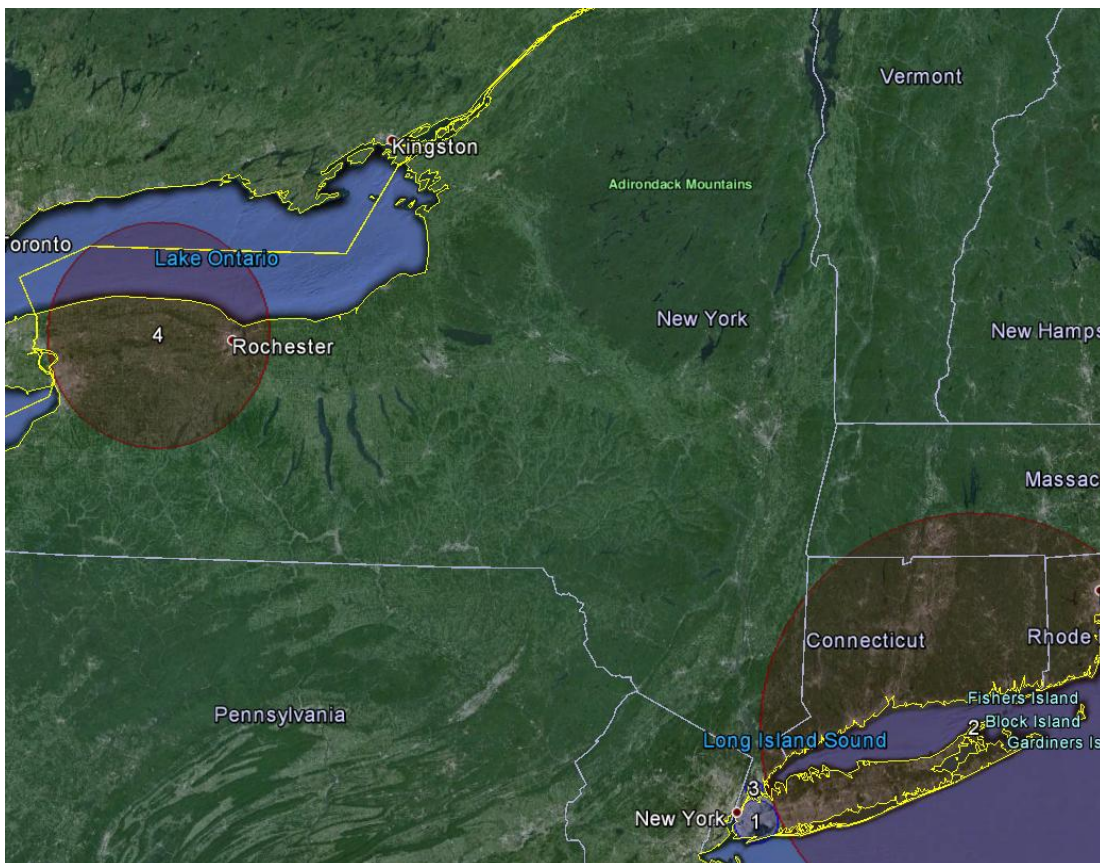
In chapter 5 it was shown how to change the maximum size of the set of potential clusters that are evaluated, and it was seen that this may lead to slightly different clusters being detected. This feature should never be used to experiment with different maxima, since an analysis with a smaller maximum size will not adjust for the multiple testing that was done when looking for larger clusters using the larger maximum. Instead, there is an advanced feature by which SaTScan will only report smaller clusters while still adjusting for the multiple testing of larger clusters. This can be used with different maxima as many times as desired while still properly adjusting for the multiple testing.

In SaTScan Tutorial #1, the maximum cluster size was set to be 25 percent of the underlying population at risk. That is, clusters were only evaluated if the expected number of breast cancer cases were less than 25 percent of the total number of breast cancer cases in the state. We will keep that maximum for the clusters evaluated and for which multiple testing is adjusted for, at the same time as we will ask SaTScan to only report clusters that are at most 10% of the population at risk.

First open the SaTScan session that was saved from Tutorial #1, as described in section 1.4. On the main *'Output'* Tab, click the *'advanced'* button in the lower right corner. You should now see the *'Spatial Output'* Tab, and at the bottom are the options for *'Maximum Reported Spatial Cluster Size'*. First click the check box titled *'Report only clusters smaller than:'*. On the next row, set the maximum at 10% of the population at risk.



Note that the 'Maximum Spatial Cluster Evaluated Analysis' tab -> 'Advanced Options' -> 'Spatial Window' Tab will still be 25%, and should not be changed. So, while cluster with up to 25% of the population at risk will be evaluated and adjusted for, only clusters with at most 10% will be reported. The results follow:



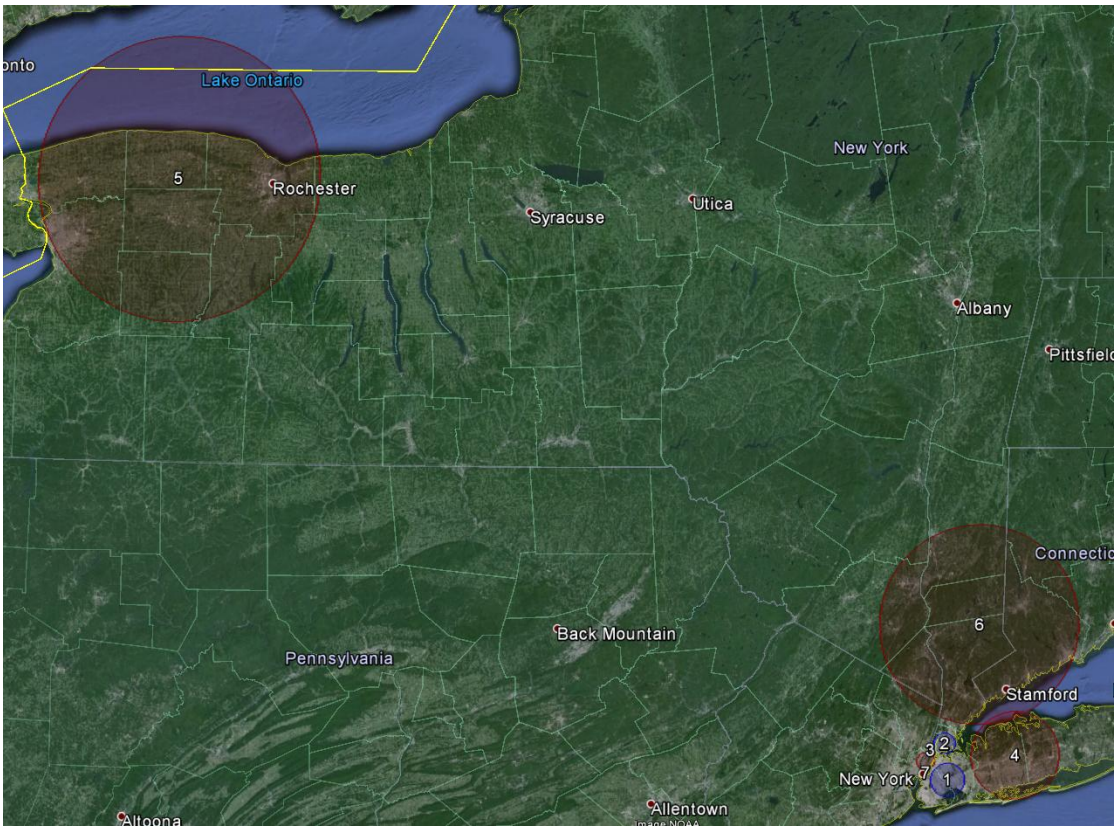
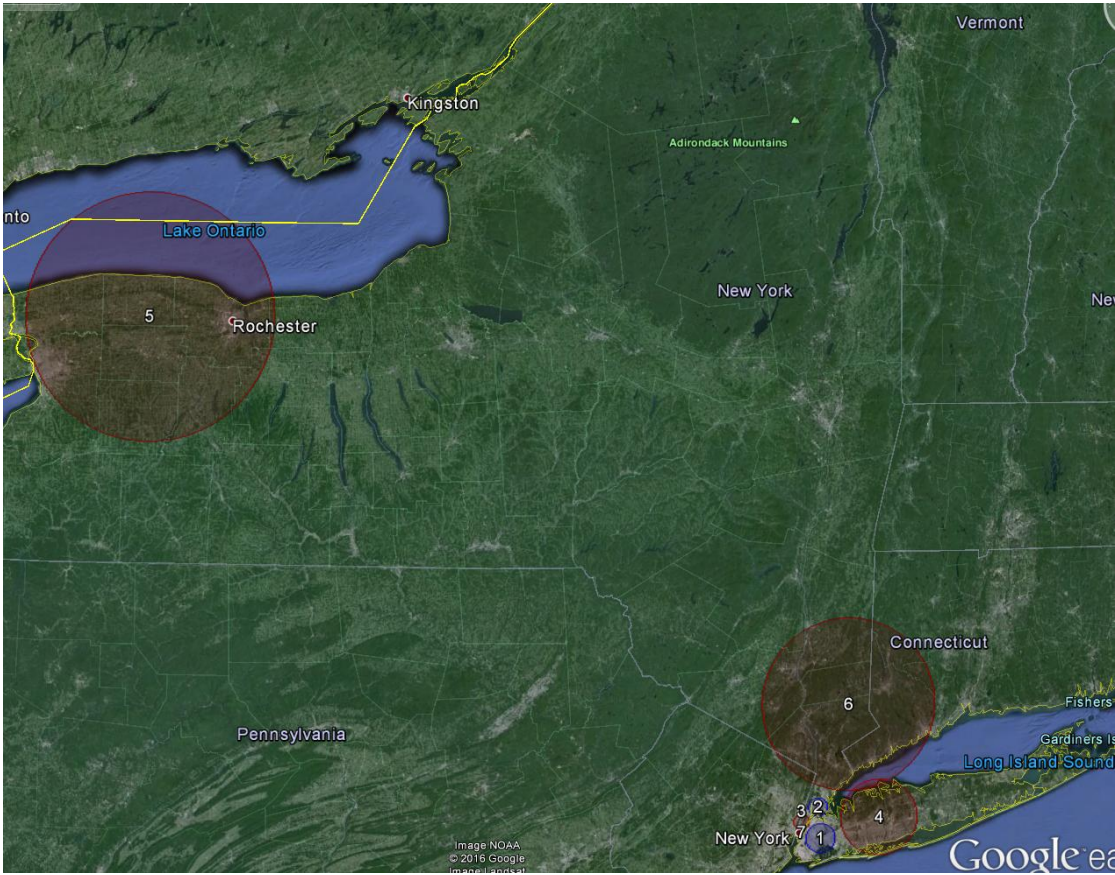


Figure 12: Breast cancer incidence clusters in New York State when the circular spatial scan statistic was run with maxima of 25% for evaluated and reported cluster (top), 10% for both evaluating and reporting clusters (middle), and 25% for evaluating and 10% for reporting clusters (bottom)

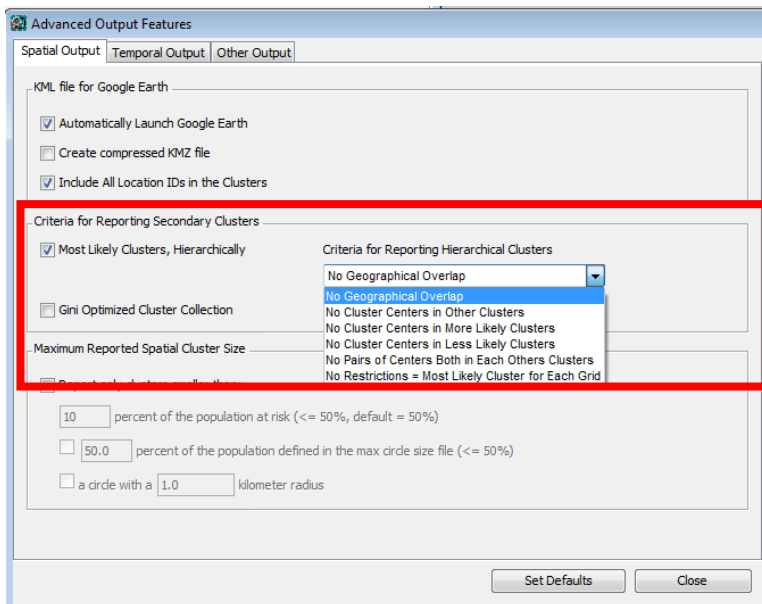
Evaluating 25% Reporting 25% (Tutorial #1)		Evaluating 10% Reporting 10% (Chapter 5)		Evaluating 25% Reporting 10%	
Cluster	p-value	Cluster	p-value	Cluster	p-value
1	< 0.0000001	1	< 0.0000001	1	< 0.0000001
2	< 0.0000001	4	< 0.0000001	4	< 0.0000001
		6	0.0000000000016	6	0.0000000000017
3	< 0.0000001	2	< 0.0000001	2	< 0.0000001
4	< 0.0000001	5	< 0.0000001	5	< 0.0000001
5	0.188	8	0.185	8	0.188

Table 10: Breast cancer incidence clusters in New York State when the circular spatial scan statistic was run with different maxima for evaluation and reporting clusters.

Figure 12 and Table 10 show the results of the new analysis evaluating clusters with at most 25% of the population, while only reporting clusters with at most 10%; as well as the results of the prior analyses. As expected, the new analysis has the exact same clusters as the Chapter 5 analysis with a 10% maximum for both evaluating and reporting clusters. Note is the difference between the p-values though. The p-value for the corresponding clusters are smaller, 0.185 as opposed to 0.188, for the latter because there is less multiple-testing done. In fact, the p-values for the new analyses are identical to the Tutorial #1 analyses, since they both evaluate the same set of clusters, so the multiple testing to adjust for is the same.

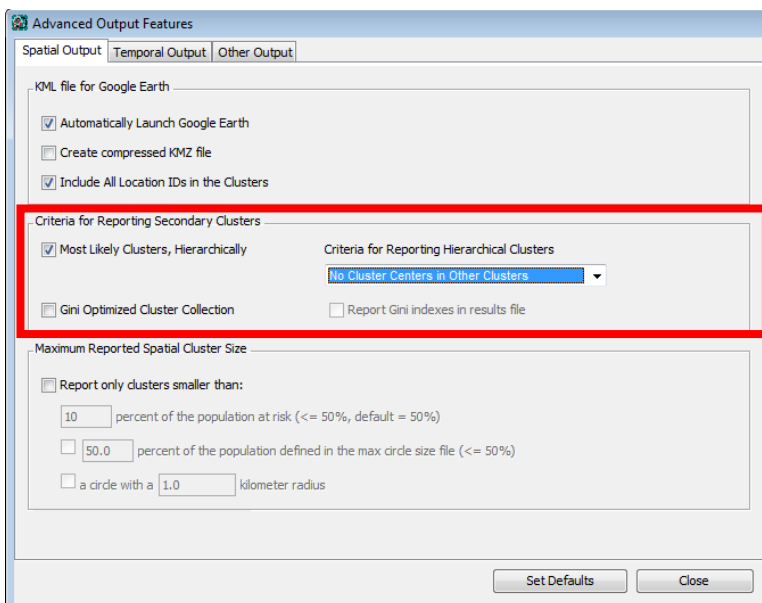
### 6.3 Report Overlapping Clusters

Reopen the SaTScan session that was saved from Tutorial #1, as described in section 1.4. In this analysis, only non-overlapping clusters were reported. It is possible to also report selected overlapping clusters. On the main 'Output Tab', click the 'Advanced' button in the lower right corner. You should now see the 'Spatial Output' Tab. As seen here, the 'Criteria for Reporting Secondary Clusters' is 'Most Likely Clusters, Hierarchically' with 'No Geographical Overlap'. By changing the drop down menu, we will be able to examine some of the other settings by which a number of overlapping clusters will be reported.



### *No Cluster Centers in Other Clusters*

To start, choose 'No Cluster Centers in Other Clusters'. This is still fairly restrictive.



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

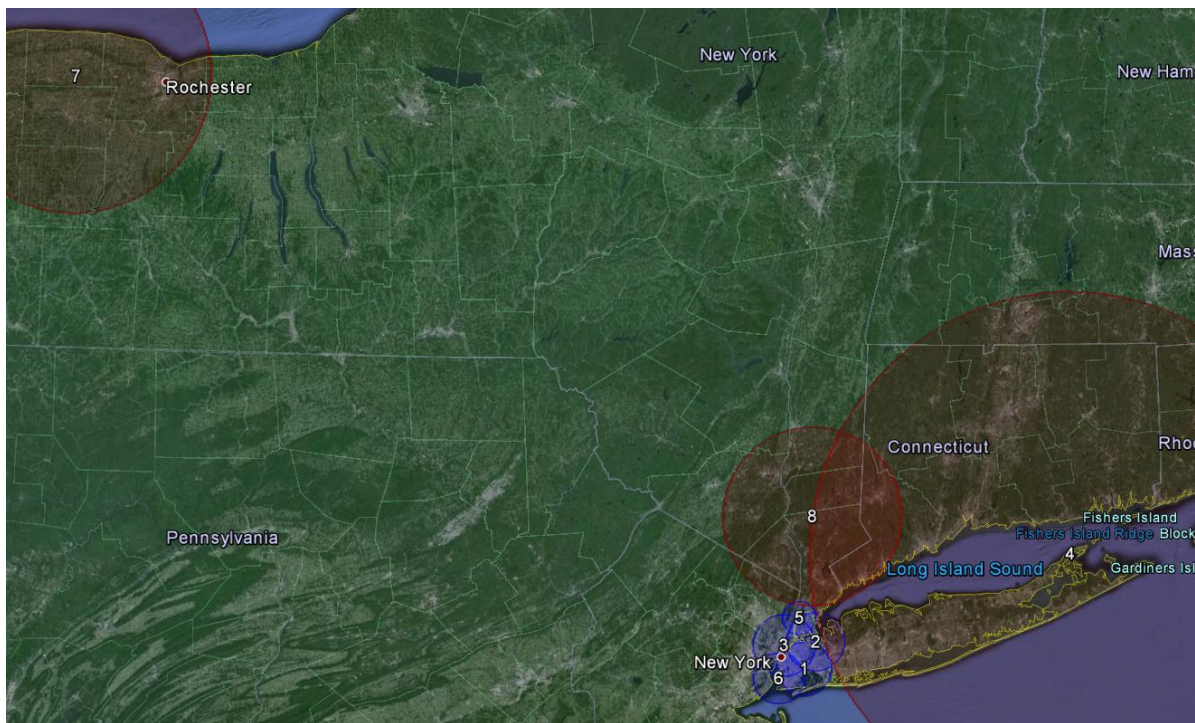


Figure 13: Reporting clusters hierarchically with No Cluster Centers in Other Clusters. A maximum cluster size of 25% of the population-at-risk was used.

No Cluster Centers in Other Clusters with at most 25% of the population at risk				No overlapping clusters at most 25% of the population at risk (figure 12, top)				No overlapping clusters at most 10% of the population at risk (figure 12, middle)			
Cluster	Radius (km)	Observed Cases	Expected Cases	Cluster	Radius (km)	Observed Cases	Expected Cases	Cluster	Radius (km)	Observed Cases	Expected Cases
1	12.82	13642	15886	1	12.82	13642	15886	1	7.80	5901	7229
2	13.91	15561	17359	No overlapping cluster				No overlapping cluster			
3	14.81	16222	18027	3	4.97	3112	3976	2	4.97	3112	3976
4	125.47	15019	13416	2	125.47	13416	15019	3	4.08	3648	2974
5	7.97	5067	6076	No overlapping cluster				4	20.41	7831	6869
6	12.18	8517	9756	No overlapping cluster				No overlapping cluster			
7	65.97	7984	7098	4	65.97	7984	7098	5	65.97	7984	7098
8	42.61	6398	5785	No overlapping cluster				6	45.93	6330	5684

Table 11: Comparison of No Cluster Centers in Other Clusters with almost 25% of the population at risk versus 25% and 10% of the population at risk

The cluster detected around Rochester in Upstate New York is the same as before, but things look different around New York City. There are two overlapping high incidence rate clusters, identical and similar to the clusters found in the prior analyses with a 25 and a 10% maximum reporting size, respectively. In New York City, the collection of overlapping low incidence clusters give a more complete picture than before, showing that almost all of the city is part of a low incidence cluster.



## No Pairs of Centers Both in Each Others Clusters

We will now try a much less restrictive option, that will generate a lot more overlapping clusters, by selecting 'No Pairs of Centers Both in Each Others Clusters'.

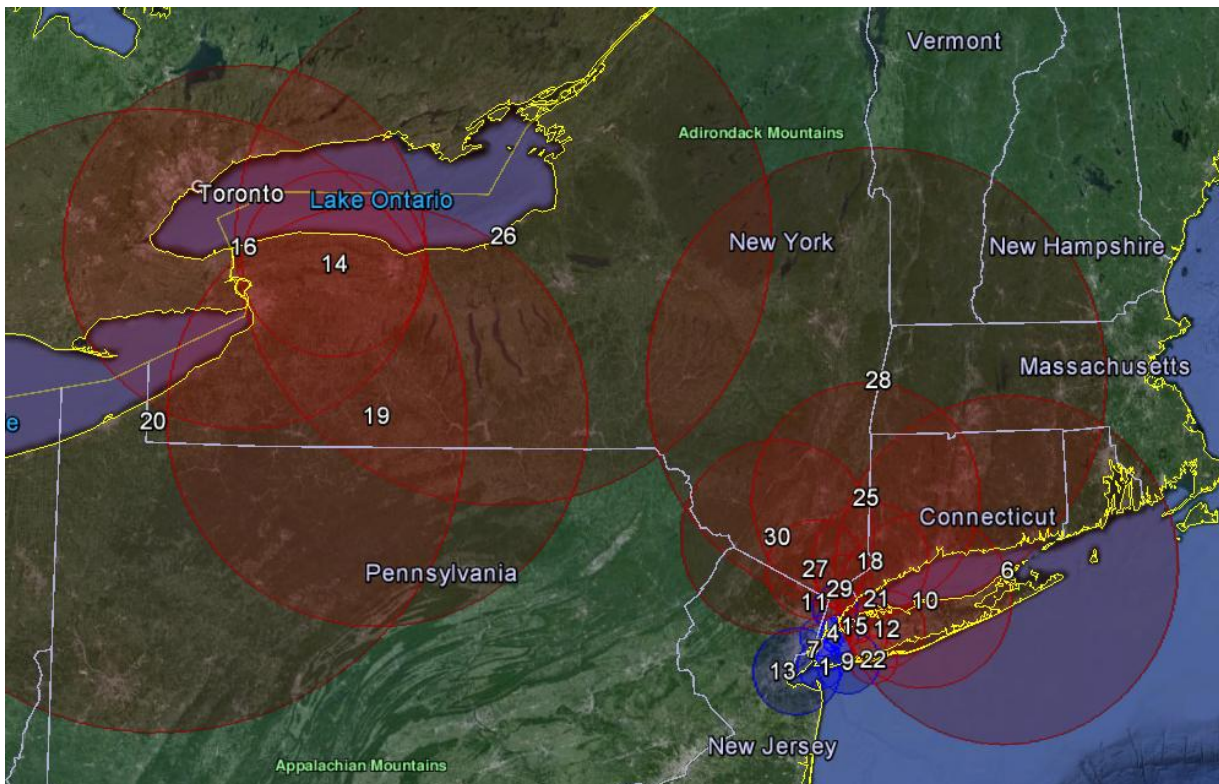
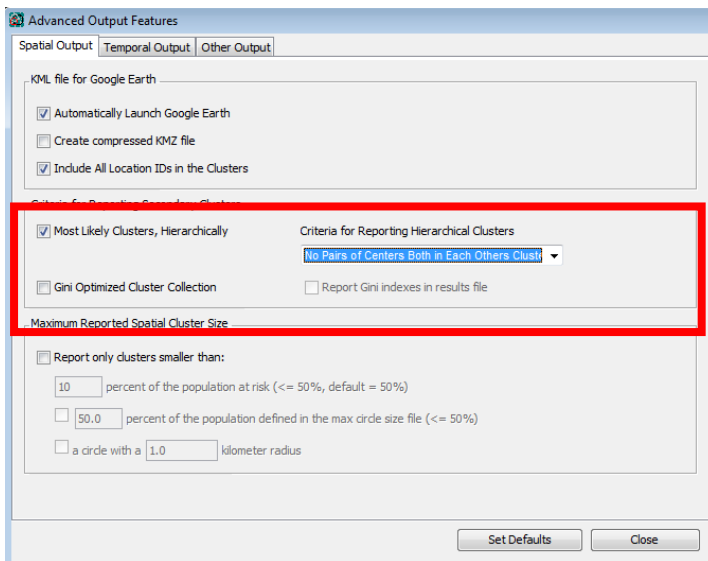


Figure 14: Reporting clusters hierarchically with No Pairs of Centers Both in Each Others Clusters size of 25% of the population-at-risk was used.

Cluster	The whole cluster			Overlap with Cluster 14			Outside Cluster #14		
	Observed Cases	Expected Cases	Observed/Expected	Observed Cases	Expected Cases	Observed/Expected	Observed Cases	Expected Cases	Observed/Expected
14	7984	7098	1.12	7984	7098	1.12	0	0	N/A
16	9745	8822	1.10	7968	7089	1.12	1777	1733	1.03
19	13509	12536	1.08	7984	7098	1.12	5525	5438	1.02
20	10898	10016	1.09	7950	7070	1.12	2948	2946	1.00
26	17662	16855	1.05	7984	7098	1.12	9678	9757	0.99

Table 12: Comparison of No Pairs of Centers Both in Each Others Clusters with 25% of the population at risk versus 25% and 10% of the population at risk

These set of clusters are much less useful. If we look in the western part of the state, the most likely cluster is #14 centred just outside Rochester. The four surrounding clusters, #16, 19, 20 and 26, cover most of the same area, but with their centroids being forced to be outside a previously reported cluster. Note that the parts of clusters #16, 19, 20 and 26 that overlap with cluster #14 have the same observed/expected as cluster #14, while the parts that do not overlap with cluster #14 have an observed/expected ratio that is close to one. Hence, the excess risk in these four clusters is almost solely driven by cluster #14 and once cluster #14 is detected, these other clusters do not provide any useful additional information. For most analyses, we do not recommend using this option.

## Chapter Seven: Gini Clusters

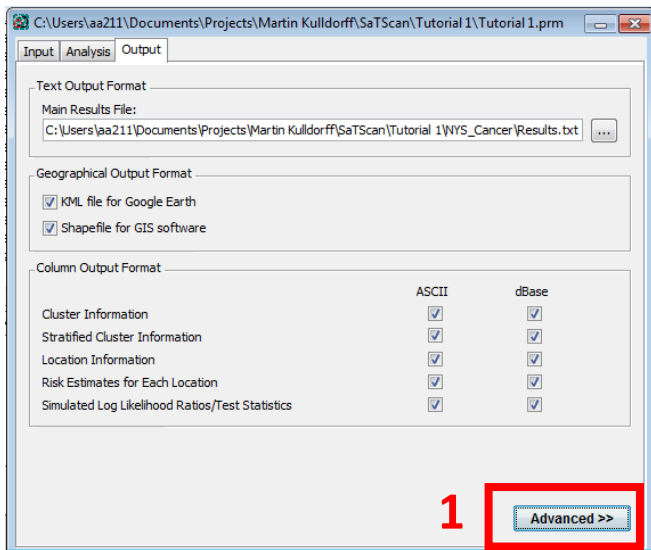
### 7.1. Background

In Chapter 6, we saw that different criteria for reporting clusters will produce different collections of non-overlapping or overlapping clusters. What is the best collection? If we report non-overlapping clusters hierarchically, some important smaller clusters may be subsumed into larger clusters and missed. On the other hand, if we only report smaller clusters, important larger clusters may be missed. If we only want to show a set of non-overlapping clusters, is it better to report a big cluster or is it better to report one or more smaller clusters that overlap with the larger one? SaTScan has a feature to determine that, using the Gini index, creating a set of non-overlapping '*Gini clusters*' (Han et al 2016).

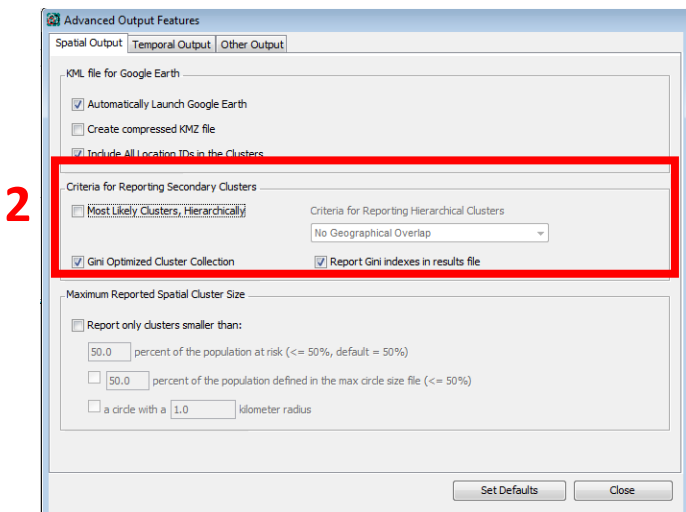
To create the collection of Gini cluster, SaTScan first defines a collection of upper limits on the reported cluster size, which in our case will be 1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20 and 25 percent of the population at risk. For each upper limit, the hierarchical no-geographical overlap cluster collection criterion is used to define a set of clusters. The Gini index is then calculated for each set of clusters, and when repeated for each upper limit, we get twelve different sets of clusters with different Gini indexes. SaTScan then picks the collection that maximizes the Gini index. This collection is called the '*Gini clusters*'. In essence, the Gini index determines if there is more evidence for one or more big cluster or multiple smaller clusters.

### 7.2 Selecting Gini Clusters

Reopen the SaTScan session that was saved from Tutorial #1, as described in section 1.4 above.



After clicking the advanced options for output, a screen like this will appear:

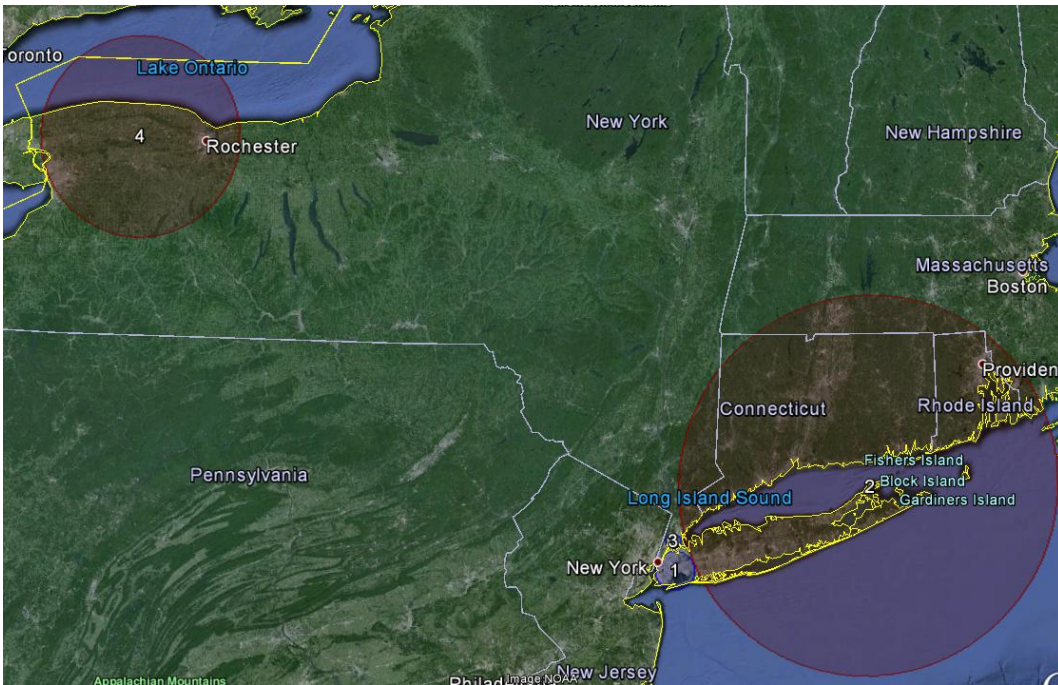
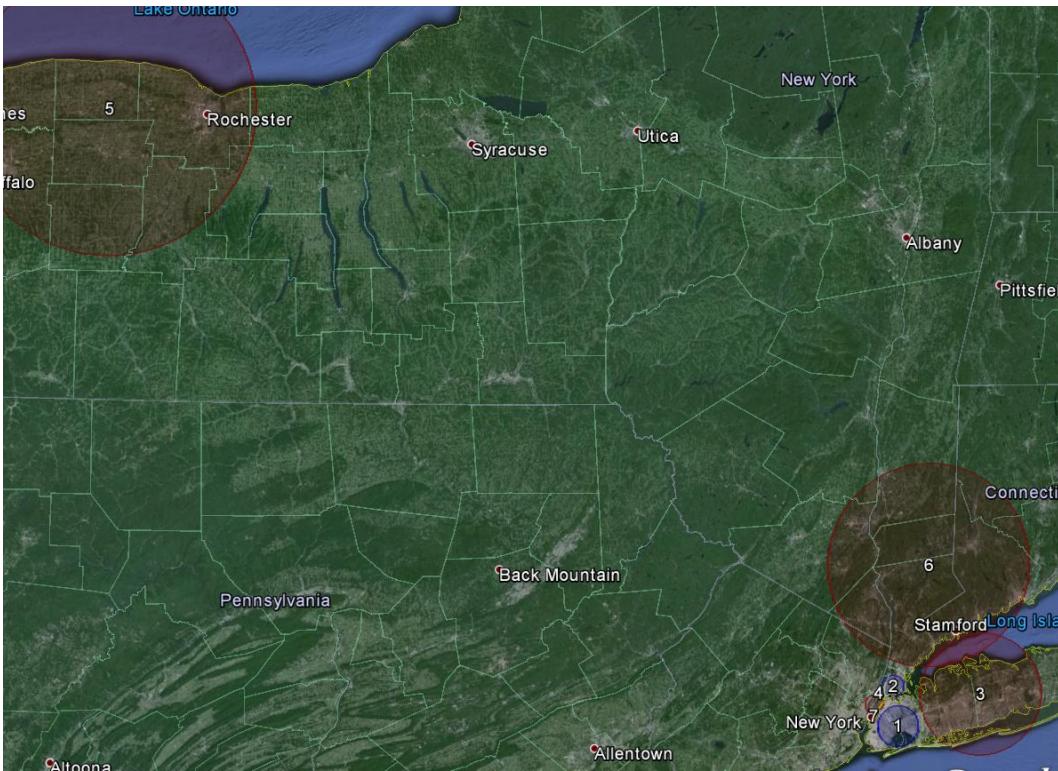


In the section '*Criteria for Reporting Secondary Clusters*', Tutorial #1 was run with '*Most Likely Clusters, Hierarchically*' reported with '*No geographical Overlap*'. As shown above, this option off and select '*Gini Optimized Cluster Collection*' instead. In addition to this, select '*Report Gini indexes in results file*'. These indexes will help us to interpret the results, as described later in this chapter.

The advanced options can now be closed and the analysis can be run.

## 7.4 Results

The results from the Gini analysis are shown in Figure 15.



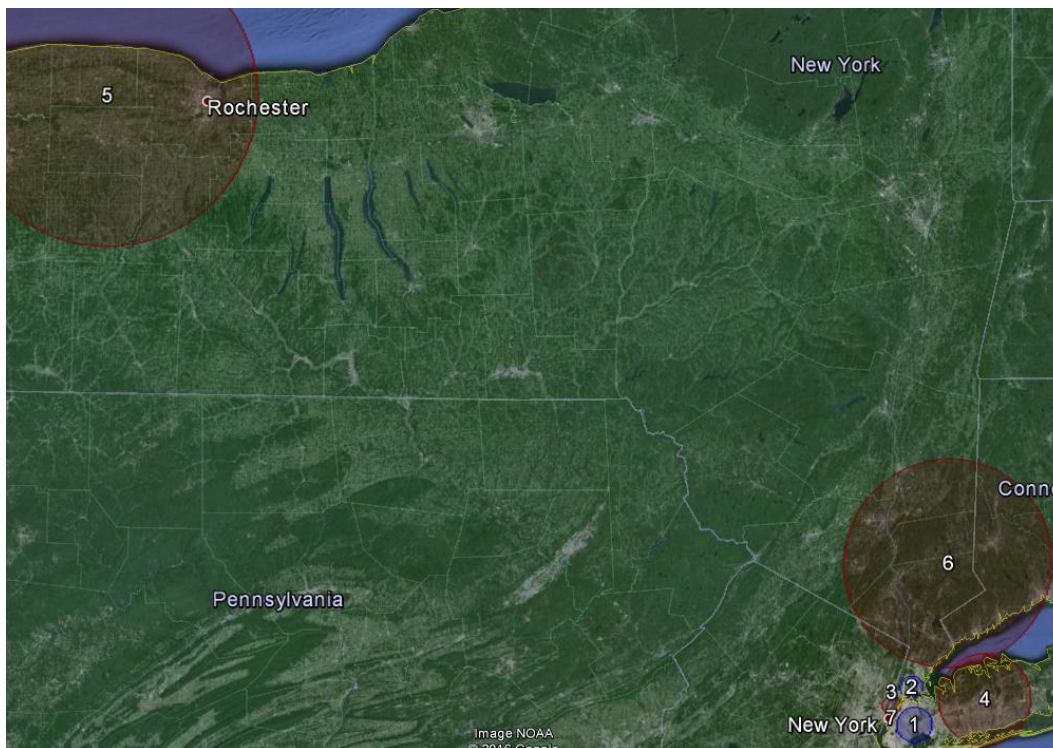


Figure 15: Breast cancer incidence clusters in New York State when the circular spatial scan statistic was run with a Gini clusters (top) versus 25% (middle) versus a 10% (bottom) maximum cluster size.

In a quick comparison of Gini cluster results it is evident that the Gini clusters resemble the analysis with the 10% maximum reporting size. A look at the Gini indexes gives further insight. These can be found at the bottom of the output file:

### Gini Indexes

-----

1 percent.....	: 0.05157
2 percent.....	: 0.06127
3 percent.....	: 0.05927
4 percent.....	: 0.06126
5 percent.....	: 0.06103
6 percent.....	: 0.06126
8 percent.....	: 0.06175
10 percent.....	: 0.06047
<b>12 percent.....</b>	<b>: 0.06218</b>
15 percent.....	: 0.05981
20 percent.....	: 0.06216
25 percent.....	: 0.05677

Gini Clusters				25% of the population at risk				10% of the population at risk			
Cluster	Radius (km)	Observed Cases	Expected Cases	Cluster	Radius (km)	Observed Cases	Expected Cases	Cluster	Radius (km)	Observed Cases	Expected Cases
1	9.29	7125	8643	1	12.82	13642	15886	1	7.80	5901	7229
7	0.71	167	265					7	0.71	167	265
2	4.97	3112	3976	3	4.97	3112	3976	2	4.97	3112	3976
3	27.73	9453	8369	2	125.47	15019	13416	4	20.41	7831	6869
6	45.93	6330	5684	No overlapping cluster				6	45.93	6330	5684
4	4.08	3648	2974	4	65.97	7984	7098	3	4.08	3648	2974
5	65.97	7984	7098					5	65.97	7984	7098

Table 13: Comparison of Gini Cluster results with 25% and 10% population at risk.

The highest Gini index (optimal gini coefficient) was for 12%. Thus only significant clusters with less than 12% of the population at risk are reported in the Gini analysis. Note that while Gini clusters #3 and #6 overlap on the map, they are actually non-overlapping clusters since they do not have any DOH regions in common and the circles only overlap over the water in Long Island where no one lives.

## References and Further Reading

This is the third in a series of SaTScan tutorials. The prior two tutorials are:

*SaTScan Tutorial #1: Purely Spatial Poisson Scan Statistic for Cancer Incidence*

*SaTScan Tutorial #2: Purely Spatial Bernoulli Analysis for Birth Defect Data*

As a complement to this tutorial, we also strongly recommend reading the [SaTScan User Guide](#), as well as scientific papers describing the theory or application of the Poisson based purely spatial scan statistic:

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

Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast cancer in northeastern United States: A geographical analysis. *American Journal of Epidemiology*, 146:161-170, 1997. [[online](#)]

Sheehan TJ, DeChello LM, Kulldorff M, Gregorio DI, Gershman S, Mroszczyk M. The geographic distribution of breast cancer incidence in Massachusetts 1988-1997, adjusted for covariates. *International Journal of Health Geographics*, 2004, 3:17. [[online](#)]

Boscoe F, Talbot T, Kulldorff M. Public domain small-area cancer incidence data for New York State, 2005-2009. *GeoSpatial Health*, 2015 [[online](#)]

Abrams A, Kleinman K, Kulldorff M. Gumbel based p-value approximations for spatial scan statistics. *International Journal of Health Geographics*, 2010 [[online](#)]

Han J, Zhu L, Kulldorff M, Hostovich S, Stinchcomb D, Tatalovich Z, Lewis D, Feuer E. Determining optimal cluster reporting sizes for spatial scan statistics. *International Journal of Health Geographics*, In Press, 2016 [[online](#)]

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:*

[Abdurrahman Abdurrob](#); [aabdurrob@partners.org](mailto:aabdurrob@partners.org)

[Martin Kulldorff](#); [kulldorff@satscan.org](mailto:kulldorff@satscan.org)

---

---



