SaTScan Tutorial #1

Purely Spatial Poisson Scan Statistic for Cancer Incidence

> Thomas Talbot and Sanjaya Kumar New York State Department of Health Martin Kulldorff Harvard Medical School November, 2014 v1.0

1. Introduction

This is a step-by-step tutorial for 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 disease clusters, and to determine if they are statistically significant. The software may also be used for similar problems in other fields such as archeology, criminology, demography, ecology, geography or zoology. A wide list of published application areas can be found in the SaTScan bibliography: <u>http://www.satscan.org/references.html</u>

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

In this first SaTScan tutorial, we will use the purely spatial scan statistic to analyze the geographical distribution of breast cancer incidence in New York State, USA, 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.

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

2. New York State 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, which is a part of the New York State Department of Health. It can be downloaded from the https://health.data.ny.gov/d/cw3n-fkji?category=Health&view_name=Cancer-Mapping-Data-2005-2009 website. In addition to the downloadable data, an interactive map of the data can be viewed on the NYSDOH web site.

Geographical Resolution: The data are provided at the level of the census block group. Block groups are relatively homogeneous statistical units and are the smallest unit for which

sample-based data are tabulated by the United States Census Bureau's American Community Survey (United States Census Bureau, 2014). In order to protect patient confidentiality, a block group needs to have a minimum of six tumors diagnosed among males and females separately. Block groups not meeting this threshold were merged with neighboring block groups. This resulted in a reduction in the number of block groups.

Expected Counts: The expected counts were calculated using the indirect standardization method, adjusted for 5-year age groups up to 85+, using the 2010 census counts for New York State. In exchange for such fine geographic detail, other aspects of the data had to be omitted to ensure patient confidentiality. Accordingly, there is no information on the age, race, ethnicity, or any other demographic characteristics of the cases.

Socio-economic Status: Along with New York State Cancer data, two separate files containing socioeconomic status (SES) data fields and block group crosswalk are also provided. Each of these files contains a common field, DOHREGION, which can be used for linking the cancer data with these files for further analysis. The SES file includes block group level information by sex and race, ethnicity, average household size, number of occupied and vacant housing units, number of persons above and below the poverty line, median household income, and number of persons with and without a high school education. The population counts included in this file are from the 2010 United States Census and the socioeconomic data is from the 2006-2010 American Community Survey.

More information about the New York State Cancer Registry and breast cancer can be found at the <u>http://www.health.ny.gov/statistics/cancer/registry/</u> and <u>http://www.health.ny.gov/diseases/cancer/breast/</u> websites, respectively.

Data Source	Links
	https://health.data.ny.gov/d/cw3n-
	fkji?category=Health&view name=Cancer-Mapping-Data-2005-
Cancer Data	2009
Population Data	http://www2.census.gov/census_2010/04-Summary_File_1/
	http://www.census.gov/acs/www/about the survey/american c
SES Data	ommunity survey and 2010 census/

Links to Original Data Sources:

Link to the Data Set for this Tutorial: The data required for this tutorial can be found on the SatScan website, at <u>http://www.satscan.org/datasets.html</u>. The name of the file is 'NYSCancer_Region.dbf'. Please download this file to your computer before you proceed to the next step.

Data Directories:

SaTScan Input Files: *InputFiles* directory (NYSDOH Cancer Data). Data is available in txt, csv, kmz, shp and MapInfo formats in the *Data* directory

SaTScan Output Files: *OuputFiles* directory Data Dictionary: *InputFiles* directory (Data_Dictionary_SatScan_Training.xls) Socio-demographic Data: *InputFiles* directory Geographical Data: *InputFiles* directory *File required for tutorial*: NYSCancer_Region.dbf

3. SaTScan Software Download and Installation

To download the free SaTScan software, please go to <u>www.satscan.org</u>. Select "download", and follow the instructions. In order to obtain the download password, you need to register, providing your name, email address, affiliation and country.

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

4. Launching the SaTScan Software

Launch the SaTScan software by using one of the following three methods. Your set-up may vary slightly on your compter.



As the next step, select "Create New Session" from the Start Window.



You should now see the input data tab, and you are ready to specify the anaylis you want to run. The SaTScan software has three main tabs for specifying the input data, analysis parameters and output formats respectively. We will go over each in turn. Each of the three main tabs also have a set of advanced tabs to define less used software options.

5. Input Data Tab

Case File:				Time Precision		
Control File:	(Bernoulli Model)			 Month Day Generic 		
Study Period Start Date: 200	Year Month Day	Year End Date: 2000	Month	Day 31	Three data fi Case fi	separate input les are required le (*.cas)
Population File:	(Poisson Model)				Popula Coordi	nates file (*.pop)
Coordinates File:				Coordinates		
Grid File:	(optional)			 Cartesian Lat/Long 		

The first of the three main tabs is used to specify the input data. For a Poisson model, which we will be using, three input files are required: for the breast cancer incident cases, for the background population at risk and for the geographical information. The three files are linked through a location ID, which represents the name or code for some geographical entety such as a state, province, county, postal code area, census tract or dwelling. In the New York State data, the locations used are different Department of Health Regions (DOHREGION), with each one represented by a set of twelve integers.

5.1 Input Tab: Case File (*.cas)

If the Case File is in the SaTScan input file format (see User Guide), it is enough to specfy its name in the Case File text box. In our case, we do not have the breast cancer cases in such a format, so we will instead use the SaTScan Import Wizard.

Case File:		ime Precision
Control File:	(Bernoulli Model)	Month Day
ick on the	button to import the ca	ase data using the SaTScan

The SaTScan import wizard can read several common file formats including *.csv, *.xlsx, *.xls, *.dbf, *.txt and *.shp. We are using a *.dbf file.

Once the import file has been selected, you must assign columns to each of the required SaTScan variables. For the case file, we first need to specify the column for the location ID, which in our case is DOHREGION. Click on the word "unassigned" that is to the right of the Location ID variable name, and then select DOHREGION. The next step is to repeat the same procedure for the number of cases. Since we are interested in breast cancer, we select OBREAST. The O in OBREAST stands for the Observed number of breast cancer cases. After finishing this tutorial, you can reapeat the analysis for a different cancer site such as bladder, bone or brain cancer. As a third step, assign YEAR as the date/time. For a purely spatial analysis, this step is optional.



SaTScan Variable	8	Source	File Variable			Clear		
ocation ID		DOHREG	ION					
Number of Cases		OBREAS	г					
Date/Time (option	al)	YEAR			E			
ovanate i (option	nai)	unasaign						
Covariate2 (option	nal)	unassign	ed					
Covariate3 (option	nal)	unassign	ed		1000		Assign Column names	fo
Covariate4 (option	nal)	unassion	ed				the location ID the	
Generated Id #	One Count #	DOHREGION	LATITUDE	LONGITUDE	OBLAD	DER	Number of Cases and	fo
ocation1	1	360010001001	42.668063	-73.734422	4		Number of cases, and	10
ocation2	1	360010001002	42.673859	-73.740663	1		Date/Time as shown o	n
ocation3	1	360010002001	42.667679	-73.751013	1		the left	
ocation4	1	360010002002	42.659958	-73.754722	0		Life left.	
ocation5	1	360010003001	42.68651	-73.807433	2			
ocation6	1	360010003002	42.670326	-73.773726	0			
ocation7	1	360010003003	42.677833	-73.761737	7			
	1.				1.			

After assigning all the variables, click on next. You are now asked to specify the file name and the directory in which you want to save the case file. You do not have to, but if you want to, you can change the name of the case input file, to something that is easier to remember. For this tutorial, we suggest that you do not change the directory, but do change the file name to NYS_BreastCancer.cas. This file is in the SaTScan case file format, with the extention *.cas. It is automatically assigned as the case file for the current analysis.



You can change the name of the case input file to something that is easier to remember, but it's not necessary.

Cancel	Previous
Click on "Change" hutton	and browse to directory for breast cancer
Click on "Change" button	and browse to directory for breast cancer
Click on "Change" button data Click Open	and browse to directory for breast cancer

5.2 Population Data

Once you are done with the case file, repeat the same process using the population data to create the population file. For population, you want to select the EBREAST variable, which is the expected number of breast cancer cases, adjusted for age.

Population Data (*.pop):

Select Populat	ion File	Tex Toront	- Maria	×	ſ
Look in:	🔒 BreastCa	ncer	- 🦸 📁 🛙	-	
Recent Items	BlockGro BlockGro BlockGro BlockGro NYS_Brea NYS_SES NYS_SES	up_Crosswalk_region.dbf up_Crosswalk_region.prj up_Crosswalk_region.shp up_Crosswalk_region.shx astCancer.cas _Data_region.dbf _Data_region.prj _Data_region.prj		Open NYSCance	r region dh
P	NVS SES	Data_region.shy rer_region.cbf		file.	
My Documents	NYSCance NYSCance NYSCance NYSCance	er_region.pry er_region.shp er_region.shx			

ropalacion mas						
C:\Temp\SaTScan	BreastCancer WYSC	ancer_region.dbf	1			
The expected f	format of the popu <date time=""> <i< td=""><td>lation file is: Population> <</td><td>Covariate 1></td><td><covariate n=""></covariate></td><td>></td><td></td></i<></date>	lation file is: Population> <	Covariate 1>	<covariate n=""></covariate>	>	
If the selected of the expected of	file is not SaTSca rder, select the 'Ne	n formatted (w ext' button to s	rhitespace delim pecify how to n	ited) or fields a ead this file.	are not in Next >	Click Next on this file Wizard screen to import population data file.
						×
File Wizard splay SaTScan Vi	ariables For:	harte	-			
File Wizard splay SaTScan Va aTScan Variable	ariables For:	Source	: File Variable	1 mar		ear
File Wizard play SaTScan Vi aTScan Variable cation ID	ariables For:	Source	e File Variable GION			ear
File Wizard splay SaTScan Va aTScan Variable cation ID ate/Time (option	ariables For:	Source DOHRE YEAR	: File Variable GION			ear
File Wizard play SaTScan Va aTScan Variable cation ID ate/Time (option pulation	ariables For:	Source DOHREI YEAR EBREAS	e File Variable GION			ear
File Wizard play SaTScan Vi aTScan Variable cation ID ate/Time (option pulation	ariables For:	Source DOHRE YEAR EBREAS	: File Variable GION			ear ASSIGN column names fo
File Wizard play SaTScan Va aTScan Variable cation ID ate/Time (option pulation wariate 1 (option ovariate 2 (option	ariables For:	Source DOHRE YEAR EBREAS unassig unassig	: File Variable GION ST ned ned			ear ASSIGN column names fo
File Wizard play SaTScan Va aTScan Variable cation ID ste/Time (option pulation wariate 1 (option wariate 2 (option wariate 3 (option	ariables For:	Source DOHREI YEAR EBREAS unassig unassig unassig	e File Variable GION ST ned ned ned			ASSIGN column names fo the location ID, for
File Wizard play SaTScan Variable cation ID te/Time (option pulation wariate 1 (option wariate 2 (option wariate 2 (option wariate 3 (option	ariables For:	Source DOHRE YEAR EBREAS unassig unassig unassig	: File Variable GION ST ned ned ned ned			ear ASSIGN column names fo the location ID, for Date/Time and for the
File Wizard play SaTScan Va aTScan Variable cation ID ste/Time (option pulation wariate 1 (option wariate 2 (option wariate 2 (option wariate 3 (option wariate 4 (option wariate 4 (option	ariables For:	Source DOHREI YEAR EBREAS unassig unassig unassig unassig unassig	e File Variable GION ST med ned ned LONGITUDE	OBLADDER		ASSIGN column names fo the location ID, for Date/Time and for the Population as shown on
File Wizard play SaTScan Variable cation ID nte/Time (option pulation wariate 1 (option pulation wariate 2 (option wariate 3 (option wariate 4 (option wariate 4 (option wariate 4 (option	ariables For:	Source DOHRE YEAR EBREAS unassig u u u u u u u u u u u u u u u u u u u	File Variable GION ST ned ned ned LONGITUDE	OBLADDER		ASSIGN column names fo the location ID, for Date/Time and for the Population as shown on the left.
File Wizard play SaTScan Variable cation ID te/Time (option pulation wariate 1 (option variate 2 (option variate 3 (option variate 4 (option	ariables For:	Source DOHRE YEAR EBREAS unassig unassig unassig LATITUDE 42.668063 42.673859	File Variable GION st ned ned ned LONGITUDE -73,734422 -73,740663	OBLADDER 4		ear ASSIGN column names fo the location ID, for Date/Time and for the Population as shown on the left.
File Wizard play SaTScan Variable cation ID tet/Time (option pulation wariate 1 (option wariate 2 (option wariate 3 (option wariate 3 (option wariate 4 (option wariate 3 (option wariate 4 (option wariate 4 (option wariate 4 (option wariate 3 (option wariate 3 (option wariate 4 (option wariate 3 (opt	ariables For:	Source DOHRE YEAR EBREAS unassig unassig unassig unassig unassig 42.668063 42.673859 42.667679	E File Variable GION 5T ned ned ned ined LONGITUDE -73.734422 -73.740663 -73.751013	OBLADDER 4 1		ear ASSIGN column names fo the location ID, for Date/Time and for the Population as shown on the left.
File Wizard splay SaTScan Vi aTScan Variable cation ID ate/Time (option pulation variate 1 (option variate 2 (option variate 3 (option variate 3 (option variate 4 (option variate 4 (option variate 4 (option variate 4 (option variate 3 (option variate 3 (option variate 3 (option variate 3 (option variate 4 (option variate 3 (option va	ariables For:	Source DOHRE YEAR EBREAS unassig unas	E File Variable GION ST ned ned ned LONGITUDE -73.734422 -73.740663 -73.751013 -73.754722	OBLADDER 4 1 1 0		ASSIGN column names fo the location ID, for Date/Time and for the Population as shown on the left.
File Wizard play SaTScan Variable cation ID ate/Time (option pulation variate 1 (option variate 2 (option variate 3 (option variate 4 (option variate 4 (option variate 4 (option variate 3 (option variate 4 (option variate 3 (option variate 3 (option variate 4 (option variate 3 (opt	ariables For:	Source DOHRE YEAR EBREAS unassig unas	E File Variable GION ST ned ned ned	OBLADDER 4 1 1 0 2		ASSIGN column names fo the location ID, for Date/Time and for the Population as shown on the left.
File Wizard splay SaTScan Variable cation ID ate/Time (option opulation ovariate 1 (option ovariate 2 (option ovariate 2 (option ovariate 3 (option ovariate 4 (option ovariate 4 (option ovariate 4 (option ovariate 4 (option ovariate 4 (option ovariate 3 (option ovariate 4 (option ovariate 4 (option ovariate 3 (option ovariate 4 (optio	ariables For:	Source POHREI YEAR EBREAS unassig 42.668063 42.65958 42.68651 42.670326	E File Variable GION ST ned ned ned 	OBLADDER 4 1 1 0 2 0 0		ASSIGN column names fo the location ID, for Date/Time and for the Population as shown on the left.
File Wizard splay SaTScan Variable cation ID ate/Time (option opulation ovariate 1 (option ovariate 2 (option ovariate 2 (option ovariate 3 (option ovariate 4 (option ovariate 4 (option ovariate 3 (option ovariate 3 (option ovariate 4 (option ovariate 3 (optio	ariables For:	Source DOHREI YEAR EBREAS UDASSIG 42.667359 42.667679 42.668051 42.67326 42.67326 42.67326 42.67326 42.67326 42.67326 42.67326 42.67326 42.67326 42.67326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.677326 42.6777326 42.6777326 42.6777326 42.6777326 42.6777326 42.6777326 42.67777326 42.6777777777777777777777777777777777777	E File Variable GION ST med med ined LONGITUDE -73.734422 -73.740663 -73.751013 -73.754722 -73.807433 -73.773726 -73.761737	OBLADDER 4 1 1 0 2 0 7		ear ASSIGN column names fo the location ID, for Date/Time and for the Population as shown on the left.
File Wizard	ariables For:	Source DOHREI YEAR EBREAS unassig unassig unassig unassig unassig 42.668063 42.673859 42.667679 42.659958 42.68651 42.670326 42.677833	E File Variable GION st med med LONGITUDE -73.734422 -73.740663 -73.751013 -73.754722 -73.807433 -73.773726 -73.761737	OBLADDER 4 1 1 0 2 0 7 7		ear ASSIGN column names fo the location ID, for Date/Time and for the Population as shown on the left.
File Wizard	ariables For: al) al) al) al) b	Source DOHRE YEAR EBREAS unassig unass	E File Variable GION ST ned ned ned LONGITUDE -73.734422 -73.740663 -73.751013 -73.751013 -73.754722 -73.807433 -73.773726 -73.761737	OBLADDER 4 1 1 0 2 0 7 7		ear ASSIGN column names fo the location ID, for Date/Time and for the Population as shown on the left.

The numbers in the population file can be either expected counts, like we have, or raw population numbers. If we had the latter, SaTScan could automatically do the age adjustment. This is done by providing SaTScan with case and population numbers stratified by age, and assigning the age variable as covariate 1 in both the case and population files. Since our expected counts are already adjusted for age, we do not have to worry about that. Instead, just click next to save the population file.

Save imported input file as	5:	
C: \Temp \SaTScan \Popula	tion.pop	Change

Click on Change and change the file name from Population.pop to NYS_BreastCancer.pop

Folder name: C:\Temp\SaTScan\BreastCancer	
	Open
Network	15

Click **Open** and type the file name (NYS_BreastCancer.pop) and then click **import**.

Save imported input file as:	
C:\Temp\SaTScan\BreastCancer\WY	S_BreastCaner.pop Change
Caus these settings and soud directly	from the second scheme second as the same baris
3 Save these setunds and read directly	v from tile source when running the analysis.
) Save these settings and read directly	y from the source when running the analysis.
) Save these setungs and read direct	y from the source when running the analysis.
Cancel	Previous

C:\Temp\SaTScan\BreastCancer\WYS_BreastCaner.pop

...

5.3 Geographical Data

We cannot do a spatial analysis without information about the spatial location of the cancer cases and the population. Using the same location IDs as for the cases and the population, we have to specify the geographical coordinates for each location ID. That is done using the coordinates file.

Coordinates File (*.geo):

Grid File:	(optional)		Open file wizard for coordinates f	ile
Select Coordin	ates File	The State State		
Look in:	BreastCancer		• 🦸 📁 🖽 •	
Recent Items	BlockGroup_Crosswalk_regio BlockGroup_Crosswalk_regio BlockGroup_Crosswalk_regio BlockGroup_Crosswalk_regio NYS_BreastCancer.cas NYS_BreastCancer.cop	n.dbf n.prj n.shp n.shx	Open the file	
Desktop My Documents	NVS_SES_Data_region.dbf NVS_SES_Data_region.prj NVS_SES_Data_region.shp NVS_SES_Data_region.shp NVS_SES_Data_region.shy NVS_SES_Data_region.dbf	-	file	
Computer	NYSCancer_region.shp NYSCancer_region.shx			
Network	File name: NYSCancer_region. Files of type: All Files	dbf	Cancel	
ile Wizard ordinates File:	Section 1			
Temp\SaTScan\B ie expected for <i>location ID</i> > <	reastCancerWYSCancer_region.dbfj mat of the coordinates file, usi <i>Latitude> <longitude></longitude></i>	ng Latitude/Longitude	coordinates is:	
the selected fil	e is not SaTScan formatted (w er, select the 'Next' button to sp	hitespace delimited) or pecify how to read thi	fields are not in s file.	o to nev
			step in File Wiz	zard

Location ID Latitude (y-axis) Longitude (x-axis)		DOHR LATIT LONG	EGION UDE ITUDE			Make selections for SatScan variables using the dropdown list under in Source File Variable
Generated Id #	DOHREGION	LATITUDE	LONGITUDE	OBLADDER	OBONE	column. You will need to
location 1	3600 1000 100 1	42.668063	-73,734422	4	0	make selection for each
location2	3600 1000 1002	42.673859	-73.740663	1	0	each SallScan Variable as
location3	360010002001	42.667679	-73.751013	1	0	shown.
location4	360010002002	42.659958	-73.754722	0	0	
location5	3600 1000 300 1	42.68651	-73.807433	2	0	(
location6	3600 1000 300 2	42.670326	-73.773726	0	0	(
location7	3600 1000 300 3	42.677833	-73.761737	7	0	
€	000040000004	10 077444		1.	10	•
# = Column is not	actually defined in	file but can be u	ised as SaTScan	variable. < Previous	Nex	t >

File Wizard	<u> </u>
Save imported input file as:	-
C:\Temp\SaTScan\BreastCancer\Coordiantes.geo	Change
Save these settings and read directly from file source when running to	the analysis.

Click on Change and change the file name from Coordinates.pop to NYS_BreastCancer.geo

	Enlder name:	Cultamol/CaTCraolBreast/Cancer	 Onen
	Foider name:	C: (Temp (Sall Scan) BreastCancer	 Open
Network	Files of type:	All Files	 Cancel

Click **Open** and type the file name (NYS_BreastCancer.geo) and then click **import**.

File Wizard	
Save imported input file as: C:\Temp\SaTScan\BreastCancer\WYS_Breast@ancer.geo Change Save these settings and read directly from file source when running the analysis.	Click Change and select file location and file name for
Cancel Cancel Import	Click Import

5.4 Information About the Input Data

After specifying the names of the input files, it is neccesary to specify some information about the input data. This is also done on the main input tab.

Time Precision (input tab)



The first piece of information to specify is the precision of the times specified in the case file. If there are no times in the case file, "none" is selected. Using integer values, generic is used for times not specified using calendar dates. In the New York cancer data that we imported, time is given as the year of the incident diagnosis, and hence we select "Year". In a purely spatial analysis, this temporal information is optional, but, SaTScan needs to know if it is in the case file in order to read that file correctly.

Study Period (Input tab)

The next step is to define the study period. This is the time period for which we have collected breast cancer cases. In this tutorial we only have data from 2009, so we select 2009 as both the start and end date. This information is important even in a purely spatial analysis, in order for the expected counts of the clusters to be correctly calculated.

Study Period									
Start Date:	Year 2009	Month 1	Day 1	End Date:	Year 2009	Month 12	Day 31		2

Coordinates (Input tab)

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



5.5 Final Check of Input Tab

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

Input Analysis Output
Case File:
C:\Temp\SaTScan\BreastCancer\WYS_BreastCancer.cas
Control File: (Bernoulli Model) 💿 Month 💿 Day
Generic 💮 Generic
_ Study Period
Year Month Day Year Month Day
Start Date: 2009 1 1 End Date: 2009 12 31
Population File: (Poisson Model)
C:\Temp\SaTScan\BreastCancer\WYS_BreastCaner.pop
Coordinates File:
C:\Temp\SaTScan\BreastCancer\WYS_BreastCancer.geo
Grid File: (optional) O Lat/Long
Advanced >>

6. SaTScan Analysis Parameter Settings

After providing SaTScan with the data needed to run an analysis, the next step is to tell SaTScan the type of analysis to perfrom. This is done on the Analysis Tab.

Analysis Tab:

6		
Inpu	Analysis Otput	
-		

The first choice is to decide whether to do the *analysis* using a purely spatial, a purely temporal or a space-time scan statistic. We are not intrested in time, so we will do a purely spatial analysis.

vpe of Analysis	Probability Model	Scan For Areas With:
etrospective Analyses:	Discrete Scan Statistics:	 High Rates Low Rates
 Purely Temporal Space-Time 	Bernoulli Space-Time Permutation Multinomial	High or Low Rates
 Spatial Variation in Temporal Trends Prospective Analyses: Purely Temporal Space-Time 	 Ordinal Exponential Normal Continuous Scan Statistics: Poisson 	Time Aggregation Units: Year Month Day Length: 1 Years
	Make selections as shown in the window.	

Next, it is necessary to choose the *probability model*. We have count data, where there is a background population from which the cases arise, and where under the null hypothesis, the cases of breast cancer are independent of each other. For such data, it is suitable to use the Poisson probability model.

Scan statistics are typically used to detect clusters of cases; that is, areas with a larger number of cases than would be expected by chance. This indicates areas where there may be a higher risk for the disease. Sometimes it is also of interest to look for areas with fewer cases than expected, where the risk of the disease is lower. In our case, let's say that we are interested in areas of either high and low risk. We then select to scan for areas with high or low rates.

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

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

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

The first of the six advanced analysis tabs is called 'Spatial Windows', which is the one we want. The default in SaTScan is to look for clusters covering up to half the population at risk, or to be more precise, half of the total expected counts. In New York State, that can be a very large area, containing almost the whole state except New York City. To avoid the detection of such large clusters, one can set a smaller maximum on the cluster size. We will pick a maximum of 25% of the population at risk, and this is specified on the top of the 'Spatial Windows' tab.

Maximum Soa	Temporal Window	Space and Time Adjustments	Inference	Border Analysis	Power Evaluation	Change
25.0 p	ercent of the populat	ion at risk (<= 50%, default =	50%)	61- (FOD)		spatial cluster size to 25%.
Ē	is a circle with a 1	.0 kilometer radius				
Include Pur	ely Temporal Cluster	s (Spatial Size = 100%)				

Note: It is inappropriate to try different values of this maximum, running multiple SaTScan analyses. Instead, pick the largest number that you could concievably be interested in, and SaTScan will look for any cluster of that size or smaller, adjusting for the multiple testing.

While we are on the advanced analaysis tabs, please go to the fourth one, which is called 'Inference'. One important parameter there is the number of Monte Carlo replications. This should be at least 999 when running an actual analysis, although it can be set to 0 or 9 for a trial run. A larger number is always better on statistical grounds, as it will increase the statistical power of the analysis. Above 999, the increase in power is very marginal. The drawback of a larger number is that the analysis takes a longer time to run. A good rule of thumb is to use 999 for large data sets that are computer intensive to run, while using 9,999 or 99,999 for smaller data sets that are quick to run. Since we have a fairly large data set, and we want this tutorial to work on both slow and fast computers, we will pick 999. Feel free to experiment with other numbers as well.

As a side note, the reason that the number of Monte Carlo replications always end 999 is that the random data sets plus the one real data set is a multiple of one thousand. This ensures that the Monte carlo based p-values have a finite number of decimals, such as 0.021

or 0.001, and that there are no p-values with an infinite number of decimals, such as 0.021729560302171....

Spatial Window Temporal Window Space and Time Adjustments Inference Borde Analysis Power Evaluation	Click on
P-Value	Inference tab.
Default	
🕐 Standard Monte Carlo	Check the
Sequential Monte Carlo Early termination cutoff: 50	maximum
Gumbel Approximation Also report Gumbel based p-values	number of
Monte Carlo Replications	number of
Maximum number of replications (0, 9, 999, or value ending in 999): 999	replications.
Prospective Surveillance	
Adjust for earlier analyses performed since:	
Iterative Scan Statistic	
Adjusting for More Likely Clusters	
Maximum number of iterations: 10 Stop when the p-value is greater than: 0.05	
	Clink Class
Set Defaults Close	CIICK Close

Once all the advanced parameter setting have been specified, click on 'Close' in the bottom right corner. If you click on 'Set Default', SaTScan will erase your choosen setting and return them to the default values.

7. Specifying SaTScan Output Options

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

Output Tab:

943			
Input Analysi: Output			Click on Outptut Tab
Type of Analysis	Probability Model	Scan For Areas With:	Click on Outplut Tab.

There are three sections on the main output tab for text, geographical and column output format.

Text Output Format		
Results File:		
Congraphical Output Format		-
deby aprilar output i ormat		
KML file for Google Earth		
Shapefile for GIS software		
Column Output Format		
Column Output Format	ASCII	dBase
Column Output Format	ASCII	dBase
Column Output Format Cluster Information Stratified Cluster Information	ASCII	dBase
Column Output Format Cluster Information Stratified Cluster Information Location Information	ASCII	dBase
Column Output Format Cluster Information Stratified Cluster Information Location Information Risk Estimates for Each Location	ASCII	dBase

Click on the icon to modify/select the location of SaTScan output file. The file will be saved as a text file and will have several important sections including a summary of the data, locaction IDs of each location included in each cluster, coordinates and radius of each cluster, population, number of cases, number of expected cases, relative risk and p-value for each cluster detected.



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

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

Advanced Output Features Spatial Output Temporal Output ut Other Output KML file for Google Earth Automatically Launch Google Earth Create compressed KMZ file Include All Location IDs in the Clusters	t		After making the selections, click on the ' Other Output " tab on top of the Advances Output Features .
Criteria for Reporting Secondary Clusters	Criteria for Reporting Hierarchical Clusters		
(E , , ,	No Geographical Overlap	•	The "Gini Optimized Cluster Collection"
Gini Optimized Cluster Collection	Report Gini indexes in results file		is deselected.

Report critical values for an observed cluster to be significant	
- Monte Carlo Rank	
Report Monte Carlo Rank	
Column Headers	Select the column headers option as
Title for Results File	the ASCII output files
	Click Close.

After you have made all the selections on the output tabs, you are ready to run a SaTScan analysis.

8. Running SaTScan

Once all the SaTScan parameters have been selected on the various tabs, it is time to run the program. This can be done in either of two ways. From the menu, first select 'Session' and then 'Execute'.

File	Session Windows He	lp
5	Execute	Ctrl+R
~	Execute Options	Ctrl+Shift+R

From Menu: Session > Execute

Alternatively, and much faster, just click on the button with the green triangle.



Once the SaTScan analysis is running, you will see a window that shows the progress being made. There is nothing in this window that you need to write down or memorize. After the analysis is done, this window will show the text-based results file. You will be able to scroll through this window to see all the results, as well as the parameter settings you used. These are the same results that are saved in the file that you specified in section 7.

Running Session	
You are running SaTScan v9.4 Alpha 4 (64-bit).	
SaTScan is free, available for download from http://www.satscan It may be used free of charge as long as proper citations are to both the SaTScan software and the underlying statistical me	n.org/. given thodology.
Reading the coordinates file Reading the population file Reading the case file Calculating the expected number of cases Constructing the circles	
Warnings/Errors:	
	Email Cancel

Sometimes SaTScan produces warnings or errors. The most common errors are problems with the input files, such as a location ID that is present in the case file but missing in the geographical coordinates file. In this tutorial, you should not get any warnings or errors if you have done everything accordining to the tutorial instructions.

9. Reading and Intepreting the Results

The SaTScan software provides the output result in various formats, some of which are optional. The main text-based results file will automatically open once the calculations are complete:

ත	C:\Temp\SaTScan\BreastCancer\OutputFiles\NYS_BreastCancer.txt	- • •
		-
	Program run on: Tue Oct 28 15:03:05 2014	
	Purely Spatial analysis	
	scanning for clusters with high or low rates using the Discrete Poisson model.	
	SUMMARY OF DATA	_
	٠ III III III III III III III III III I	
	Warnings/Errors:	
	Location 360850154001 2009/7/1	*
	Location 360850223003, 2009/7/1	
	Location 361190015034, 2009/7/1	
	Location 361190056001, 2009/7/1	-
	 III 	•
	E	imail Close

The first thing to do is to check for any warnings or errors. These are designed to help the user find any data problems that may exist in the input files as well as unintended parameter settings.

At the end of the main results file, there is a list of the parameter settings used for the analysis. For example, it states how many Monte Carlo replications were used. In this way, you can always go back and check the parameters you used for a particular analysis. At the very end, there is also information about when the SaTScan program was run, and how long it took to run it.

C:\Temp\SaTScan\BreastCancer\OutputFiles\NYS_BreastCancer.txt	
Inference P-Value Reporting - Default Combination	
Number of Replications : 999	
Rejusting for more bixery crusters . No	
Border Inslusia	
Bolder Analysis	
Report Oliveira's F : No	
Spatial Output	
Automatically Launch Google Earth : Yes	
Compress KML File into KMZ File : No	
Include All Location IDs in the Clusters : Yes	
Cluster Location Threshold - Separate KML : 1000	
Report Hierarchical Clusters : Yes	
Criteria for Reporting Secondary Clusters : No Geographical Overlap	
Report Gini Optimized Cluster Collection : No	
Restrict Reporting to Smaller Clusters : No	
Other Output	
Report Critical Values : No	
Report Monte Carlo Rank : No	
Frint ASCII Column Headers : Yes	
Run Options	
Processer Usage : All Available Proccessors	
Suppress Warnings : No	
Logging Analysis : Yes	
Program completed : Tue Oct 28 15:04:31 2014	
Total Running Time : 1 minute 26 seconds	_
Processor Usage : 8 processors	-
· · · · · · · · · · · · · · · · · · ·	+

The list of parameter settings also includes the names of the optional results files that you requested, if any. These optional files contain the results in column and row format, for easy incorporation and further analysis using GIS and other software products.

Final output files:

Organize 🔻	Include in library	Burn New folder		
^] 👢	Name	Date modified	Туре	Size
La E	S cluster1_locations.kml	10/28/2014 3:04 PM	KML File	665 KB
Jan H	cluster2_locations.kml	10/28/2014 3:04 PM	KML File	548 KB
	cluster3_locations.kml	10/28/2014 3:04 PM	KML File	171 KB
. .	cluster4_locations.kml	10/28/2014 3:04 PM	KML File	298 KB
	cluster5_locations.kml	10/28/2014 3:04 PM	KML File	106 KB
.	cluster6_locations.kml	10/28/2014 3:04 PM	KML File	9 KB
	cluster7_locations.kml	10/28/2014 3:04 PM	KML File	5 KB
	NYS_BreastCancer.col.dbf	10/28/2014 3:04 PM	Microsoft Visual F	3 KB
	NYS_BreastCancer.col.prj	10/28/2014 3:04 PM	Text Document	1 KB
	NYS_BreastCancer.col.shp	10/28/2014 3:04 PM	arcmap	40 KB
<u></u>	NYS_BreastCancer.col.shx	10/28/2014 3:04 PM	SHX File	1 KB
	NYS_BreastCancer.col.txt	10/28/2014 3:04 PM	Text Document	3 KB
	NYS_BreastCancer.gis.dbf	10/28/2014 3:04 PM	Microsoft Visual F	1,772 KB
	NYS_BreastCancer.gis.prj	10/28/2014 3:04 PM	Text Document	1 KB
	NYS_BreastCancer.gis.shp	10/28/2014 3:04 PM	arcmap	224 KB
	NYS_BreastCancer.gis.shx	10/28/2014 3:04 PM	SHX File	64 KB
	NYS_BreastCancer.gis.txt	10/28/2014 3:04 PM	Text Document	1,884 KB
	NYS_BreastCancer.kml	10/28/2014 3:04 PM	KML File	69 KB
	NYS_BreastCancer.IIr.dbf	10/28/2014 3:04 PM	Microsoft Visual F	20 KB
J S	NYS_BreastCancer.IIr.txt	10/28/2014 3:04 PM	Text Document	22 KB
	NYS_BreastCancer.rr.dbf	10/28/2014 3:04 PM	Microsoft Visual F	1,448 KB
	NYS_BreastCancer.rr.txt	10/28/2014 3:04 PM	Text Document	1,542 KB
	NYS_BreastCancer.sci.dbf	10/28/2014 3:04 PM	Microsoft Visual F	2 KB
	NYS_BreastCancer.sci.txt	10/28/2014 3:04 PM	Text Document	2 KB
	NVS_BreastCancer.txt	10/28/2014 3:04 PM	Text Document	163 KB

In order to depict the detected clusters on a map, SaTScan produces an optional geographical shapefile and/or a KML file. SaTScan will automatically open the KML file if you have Google Earth installed on your computer:



10. Some Common Questions

*How can I read the *cas, *.pop and *.geo files?*

You will need to associate the *.cas, *.pop and *.geo files to the **Notepad** program.

	Open with	
	Choose the program you want to use to open this file: File: import1455936611134463579.pop	
Windows	Adobe Acrobat 9.0 Adobe Reader Adobe Systems Incorporated	ń
Windows can't open this file:	Microsoft Corporation Diffunes	
File: import1455916611134463579.pop	Mcrosoft Excel Mcrosoft Office 2033 Mcrosoft Corporation	
To open this file, Windows needs to know what program you want to use to open it. Windows can go online to look it up automatically, or you can manually	W Mcrosoft Word	
select from a list of programs that are installed on your computer.	Notepad Microsoft Carpelation	
What do you want to do?		-
Use the Web service to find the correct program	Type a description that you want to use for this kind of file:	
Select a program from a list of installed programs	Aways use the selected program to open this kind of file	rowse
OK Cancel	If the program you want is not in the list or on your computer, you can <u>look for the appropriate program</u> Web-	amon the Cancel

import1455916611134463579.pc	p - Notepad
File Edit Format View Help	
360010001001 2009 3.56308 360010001002 2009 3.32389 360010002001 2009 6.70407	

Does SaTScan utilize all the processors on my computer for optimal speed?

Yes, SaTScan uses a paralell computing algorithm for the most computer intensive part of the calculations. You can control the processing speed by selecting the number of processors to use for the SaTScan analysis:

ile Session Windows Help	Barallel Brossessor
Execute Ctrl+R Execute Options Ctrl+Shift+R	Image: all available processors Ok Image: all available processors Ok Image: all available processors Cancel Image: all available processors Cancel Image: all available processors Cancel Image: all available processors Image: all available processors Image: all available processors Cancel Image: all available processors Image: all available processors Image: all available proce

How can I save my parameter settings?

To save time when you run the same program next time, you can save the current SaTScan session with all its parameter settings, by clicking on "File" and then "Save Session" or "Save Session As":



How can I open a saved session?

When you start SaTScan, you will be asked if you want to open an already saved session.



Alternatively, you can select "File" followed by "Open Session File" or "Reopen Session File".

Can SaTScan create kml and shp output files if my input files are in a Cartesian coordinate system?

No, SaTScan must have latitude and longitude coordinates in the input file in order to create shp and kml files.

Can SaTScan test for non-circular shaped clusters?

Yes, if the coordinate system is Cartesian, it is possible to use ellipses instead of circles. This is an advanced option on the Windows Tab.

11. References and Further Reading

This is the first in a series of SaTScan tutorials. Subsequent tutorials cover space-time scan statistics; different probability models and different types of diseases.

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

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

General Statistical Theory, Poisson Probability Model

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

Application to Breast Cancer Data with Covariate Adjustments

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]

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

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

Thomas Talbot; New York State Department of Health <u>Thomas.Talbot@health.ny.gov</u>

Sanjaya Kumar; New York State Department of Health Sanjaya.Kumar@health.ny.gov

Martin Kulldorff; Harvard Medical School martin kulldorff@hms.harvard.edu

NOTES