Spatial Cluster Detection Tutorial using SaTScan:
A Training Module for the CDC/ATSDR Guidelines for Examining Unusual Patterns of Cancer and Environmental Concerns

Table of Contents

Background

This tutorial is a **quick-start guide** to help you learn how to use SaTScan™ to **detect and explore unusual patterns in cancer data**. It's designed for **self-paced learning** but can also be used for a **classroom or webinar training**. We recommend using this tutorial alongside the SaTScan User Guide, which covers more features and options than we can include here.

This tutorial was created using **SaTScan version 10.3.2 for Windows.** These steps may look a little different if you are using a different version or operating system, but they should still be close enough to follow along.

We use sample data of **prostate cancer cases in men aged 18 years or older** from **Pennsylvania census tracts** for **2010 to 2019** as our main dataset for this tutorial. We also use population data from the **U.S. Census Bureau** to get the total male 18 years or older population in each tract. These data are all provided in the **PAcancerUpdated.xlsx** file in the tutorial folder.

Some data are **missing** (i.e., **suppressed**) to help maintain the privacy of individuals with these conditions. We have provided a separate file (**PAmiss.csv**) located in the tutorial resources to indicate tracts with missing data. We will provide step by step instructions on how to handle missing data using SaTScan during this tutorial.

The **PAcancerUpdated.xlsx** file also includes the latitude and longitude coordinates for each tract's geometric centroid as well as estimates of how many prostate cancer cases that would be expected in each tract if we adjust for the age distribution in the state population. These estimates are based on the **standardized incidence ratio (SIR)** approach. We will use these estimates in the "**Bonus Tutorial"** section at the end, to run an **age-adjusted analysis** and compare the results to our **unadjusted analysis** results.

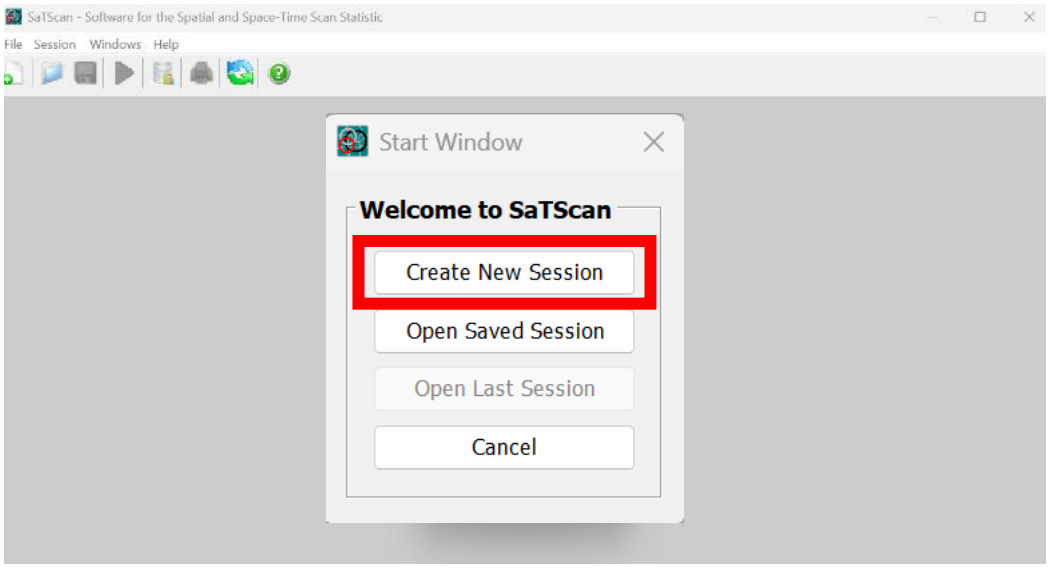
The analyses to follow are only intended as a demonstration of how to use SaTScan for cancer spatial cluster detection investigations and the sample results and findings presented as part of this tutorial should not be interpreted as real-world conclusions.

Launch SaTScan

1. Launch the SaTScan software by double-clicking on the **SaTScan shortcut** on your Desktop.

***Note**: If you do not see the SaTScan app icon, search for the application on your computer or navigate to the SaTScan folder where the software was downloaded and open the SaTScan.exe file.*

2. Once the start window appears, select the "**Create New Session"** button.

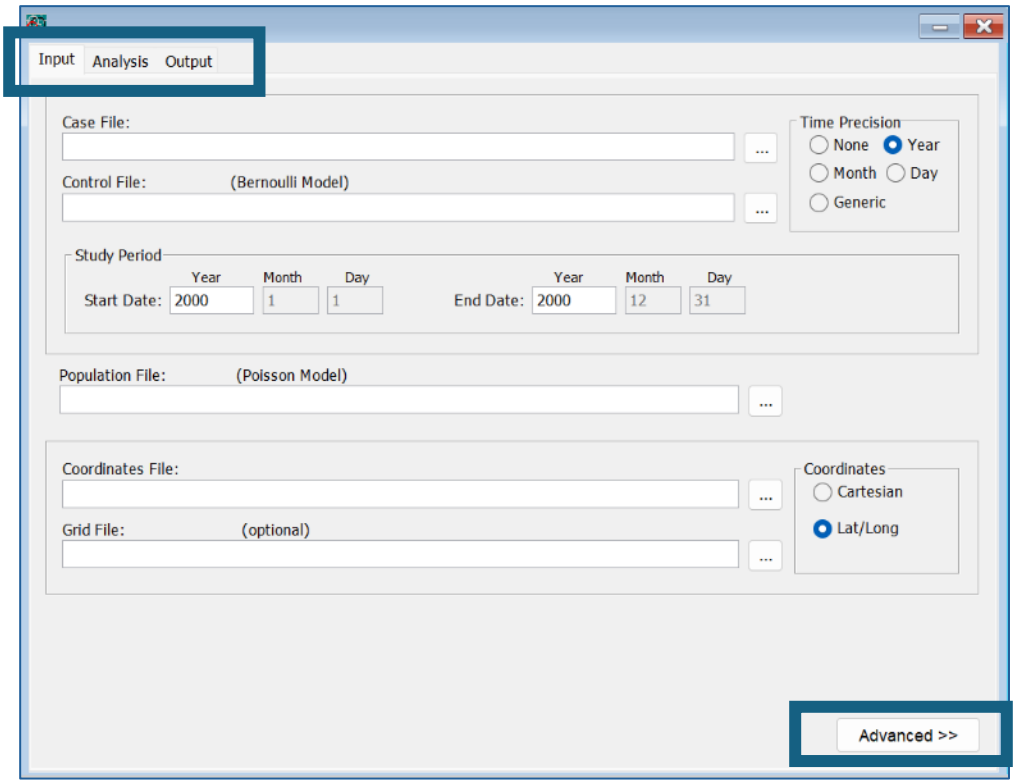


Import Data

SaTScan has 3 main tabs for specifying analysis features:

- **Input tab** - for adding all the data for your analysis
- **Analysis tab** – to specify the type of analysis you're conducting and to set the corresponding parameters for that analysis
- **Output tab** – to specify the format of the results of your analysis.

Each of these tabs also has a set of **Advanced options** (in the bottom right) to allow you to set more specialized settings for your analysis.



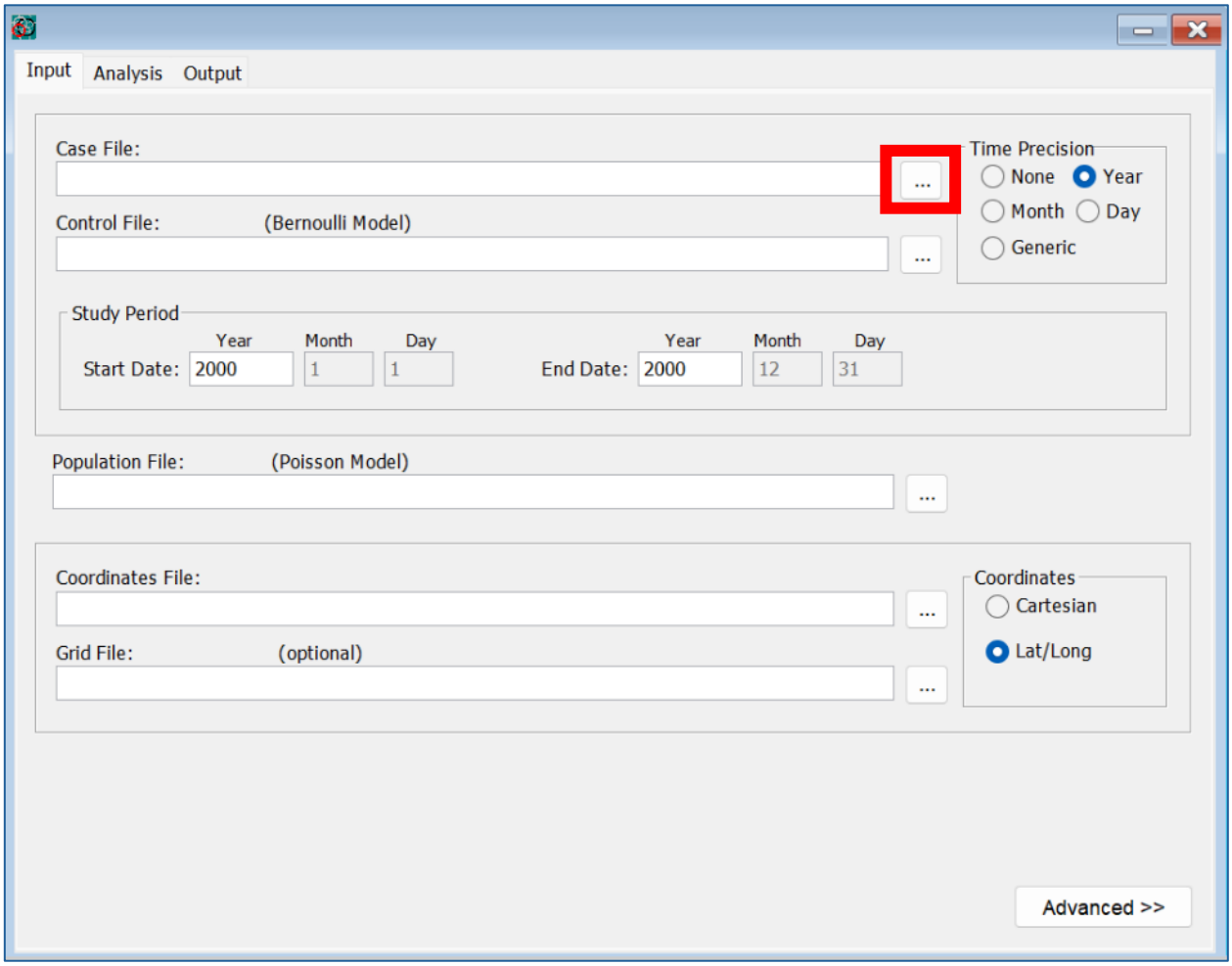
There are a few ways to add data into SaTScan, including using the SaTScan Import Wizard. As we will be running a Poisson Model for this analysis, we will need to import a **Case File**, **Population File**, and **Coordinates File**.

***Note**: Each of the files we import are automatically created and saved to a directory of your choosing, so you can easily come back to this analysis later and directly import these files, without having to go through all the SaTScan Import Wizard steps again.*

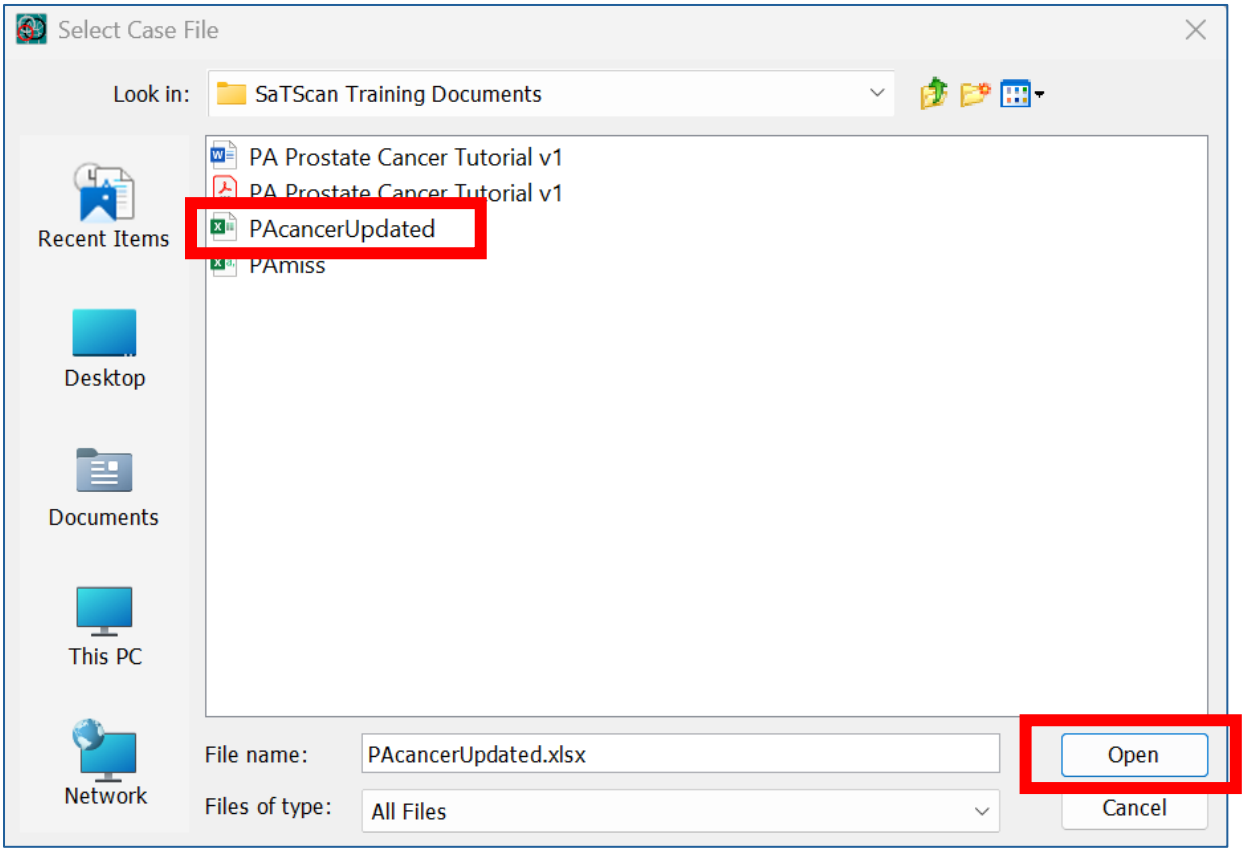
Case Data

1. Open the **SaTScan Import Wizard** by clicking on the "**ellipses (…)"** button next to the **Case File** line on the **Input** tab.

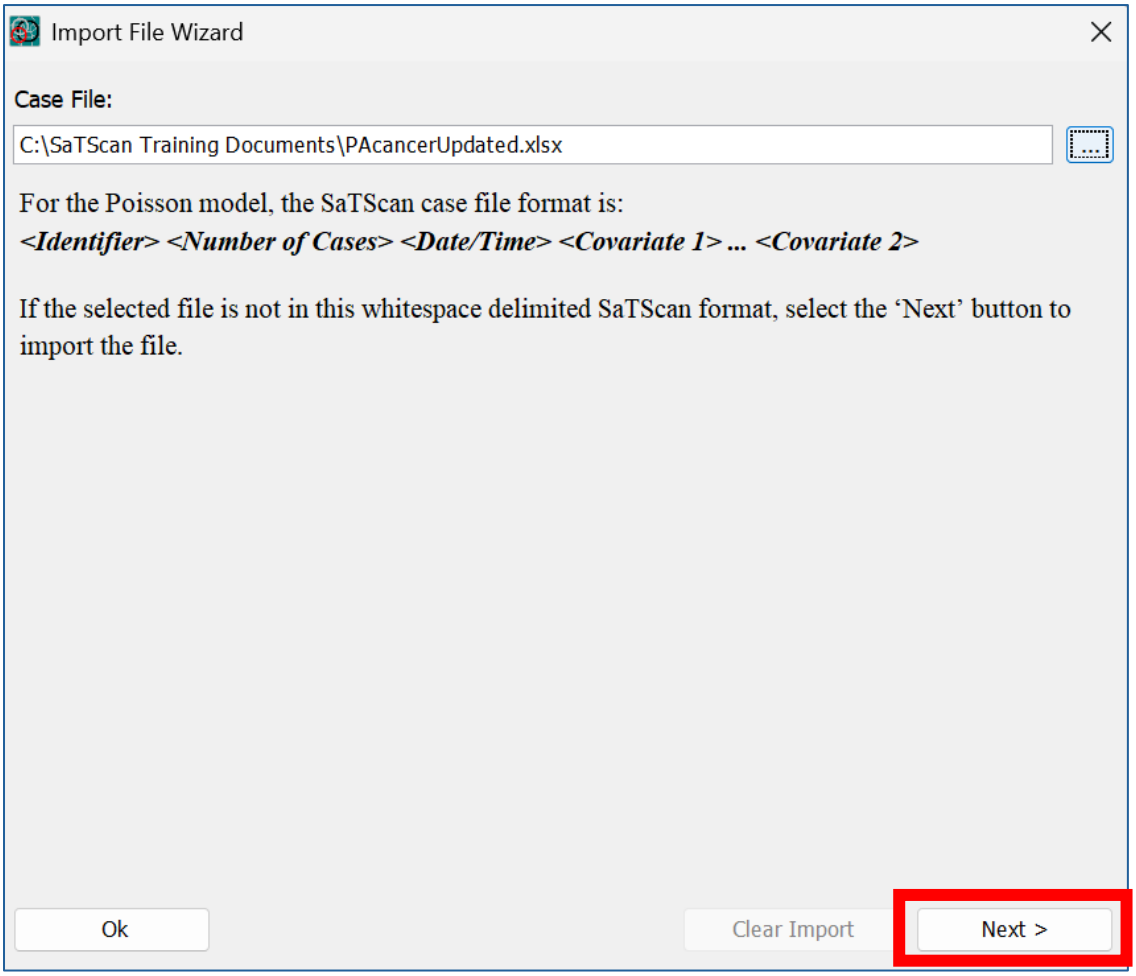
   Note: *Some parameters may already be selected by default. For example, we can ignore the Study Period section and Time Precision, as we are not performing a space-time analysis.*



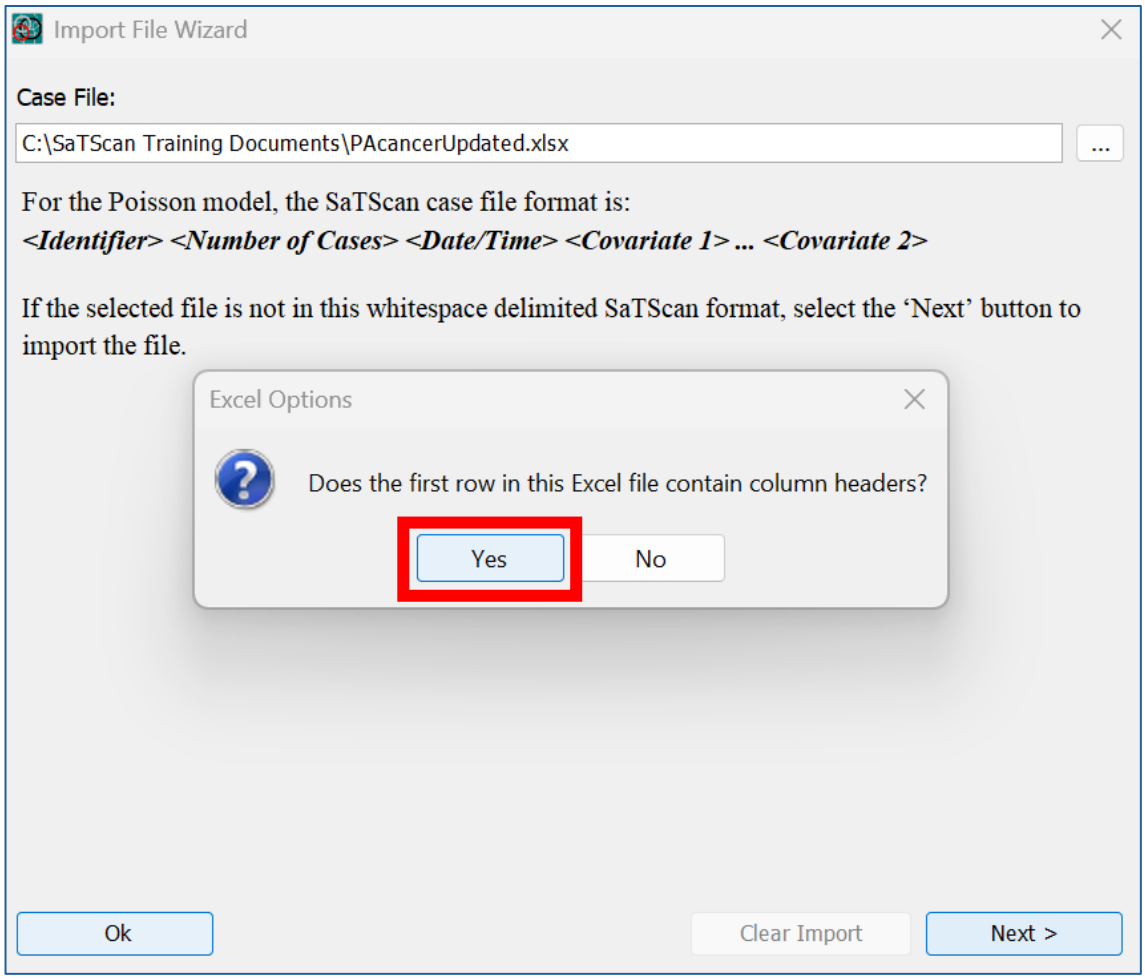
2. Navigate to where your tutorial data folder is stored, select the file called "**PAcancerUpdated.xlsx"** from the folder, and use the "**Open"** button (in the bottom right) to select it as your **Case File**.



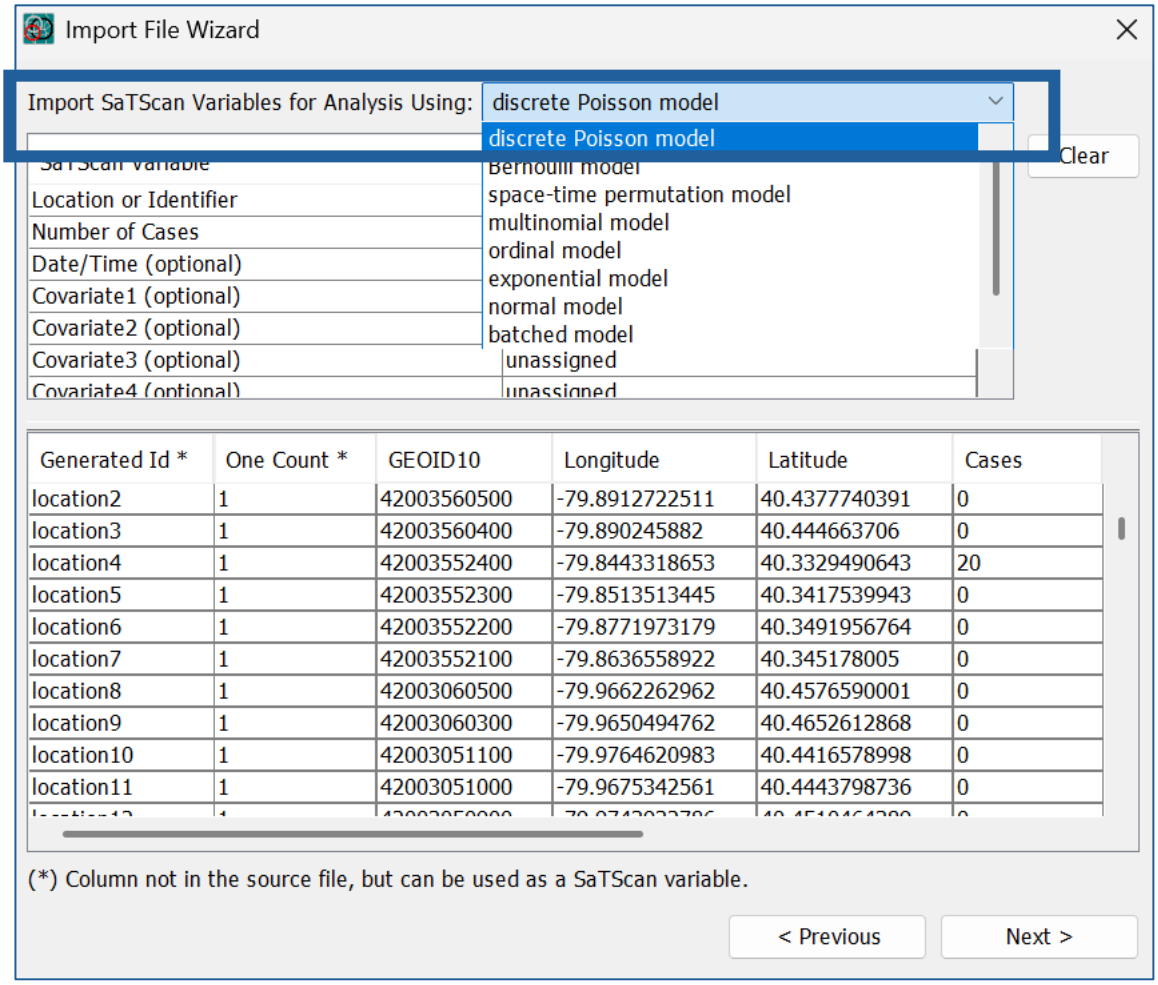
3. The window below will appear to let you know what the file format should look like. Select the "**Next**" button to continue.



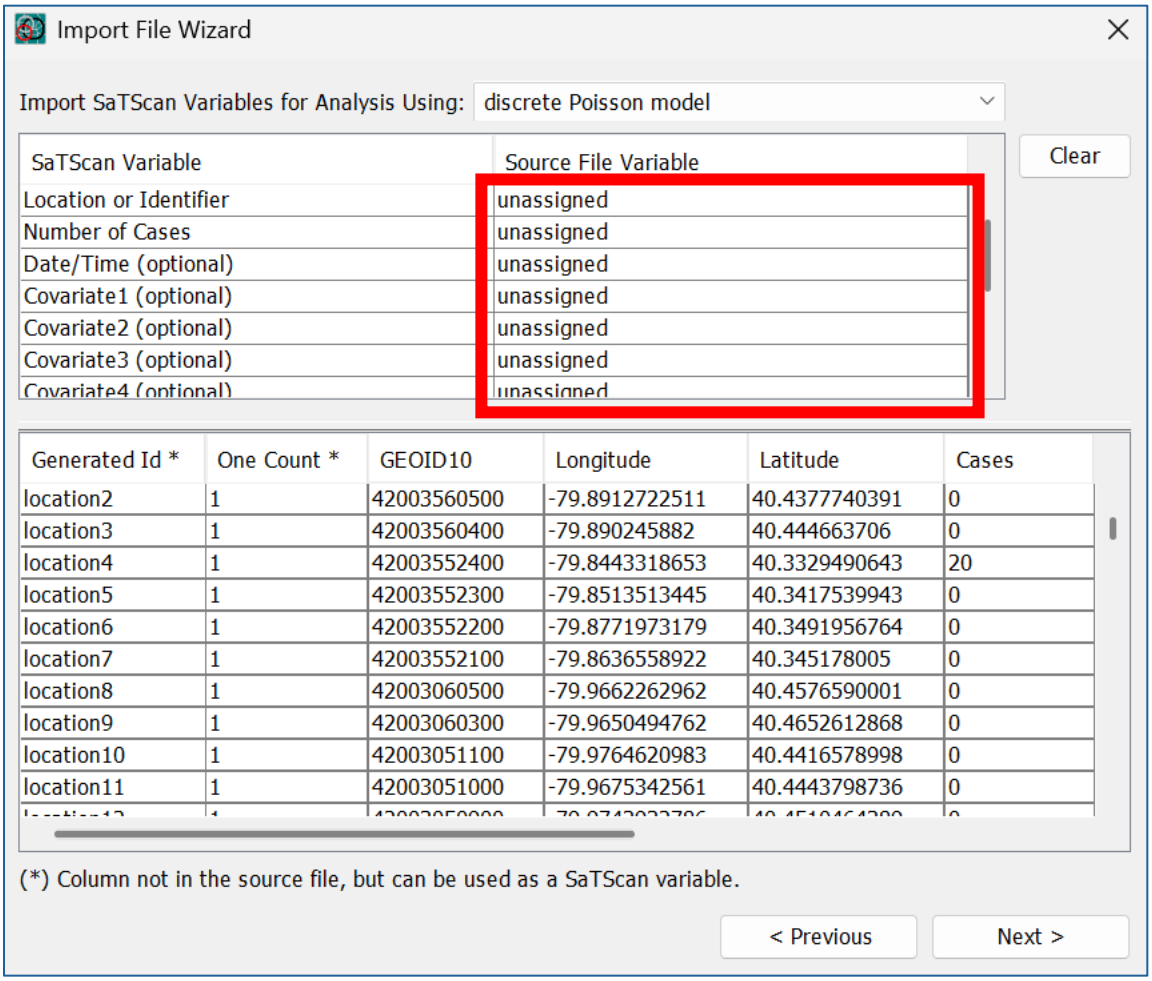
4. For Excel files (like this one) you will then get a pop-up window that asks if your file has headers. Select "**Yes**" as this file does have headers.



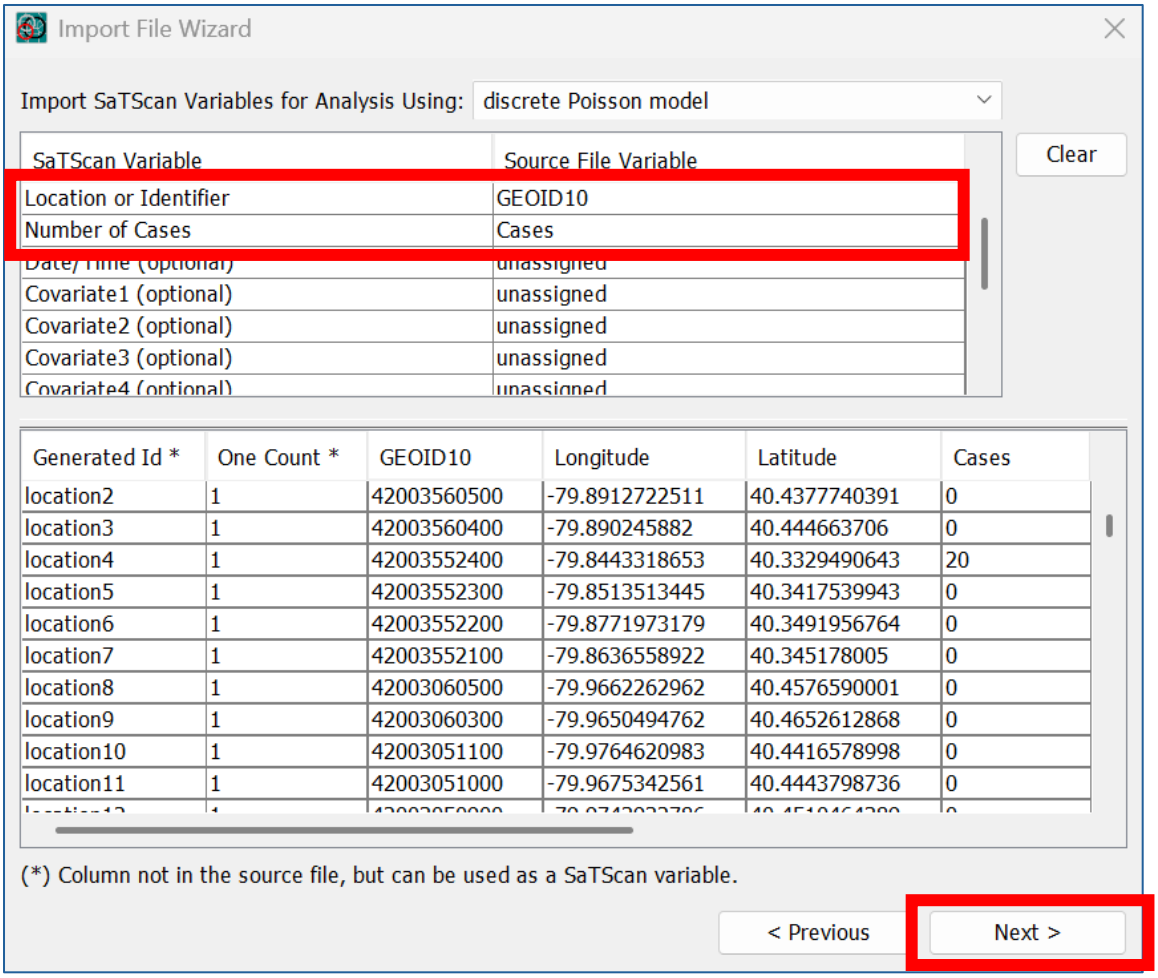
5. For this analysis, we will be using the "**discrete Poisson model**" which appears by default. However, you can always select other options under the "**Import SaTScan Variables for Analysis Using:**" drop-down menu.



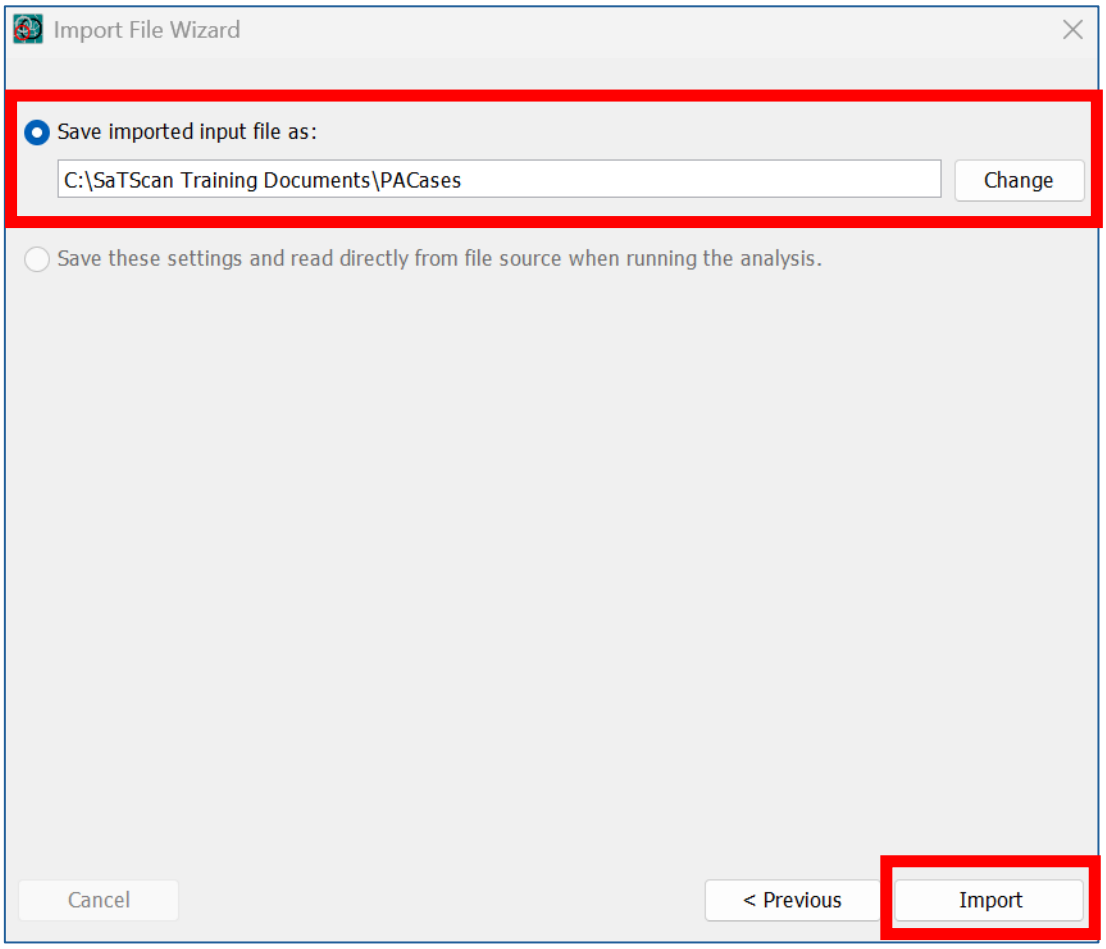
6. Assign each required **SaTScan Variable** we will be using for this analysis to a **Source File Variable**, by clicking on the "**unassigned**" box next to each variable in the **SaTScan Variable** list and choosing the appropriate variable from the drop-down menu.



7. For this analysis, assign **[Location ID]** to **[GEOID10]** and **[Number of Cases]** to the **[Cases]** variable in the Source File. Leave the remaining SaTScan variables unassigned. Select the "**Next**" button to move to the next screen.

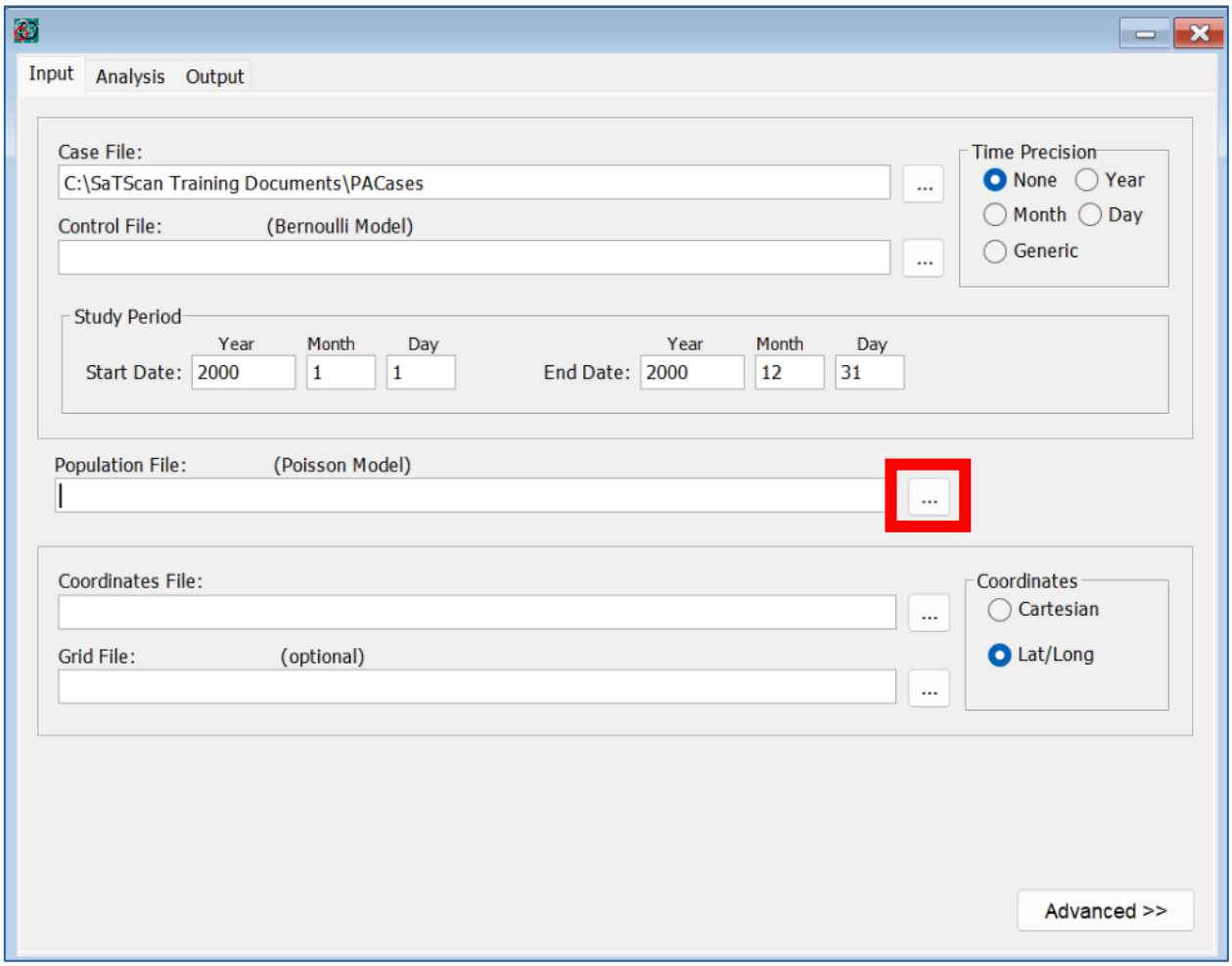


8. Import the variables you assigned (i.e., **[Location ID]** and **[Number of Cases]**) into a new file that SaTScan will read the case information from. Update the default file location and name in the address bar or by selecting the "**Change**" button to choose a new location. Update the file name in the address bar to **"PACases"** for this tutorial. Select the "**Import**" button in the bottom right when finished.

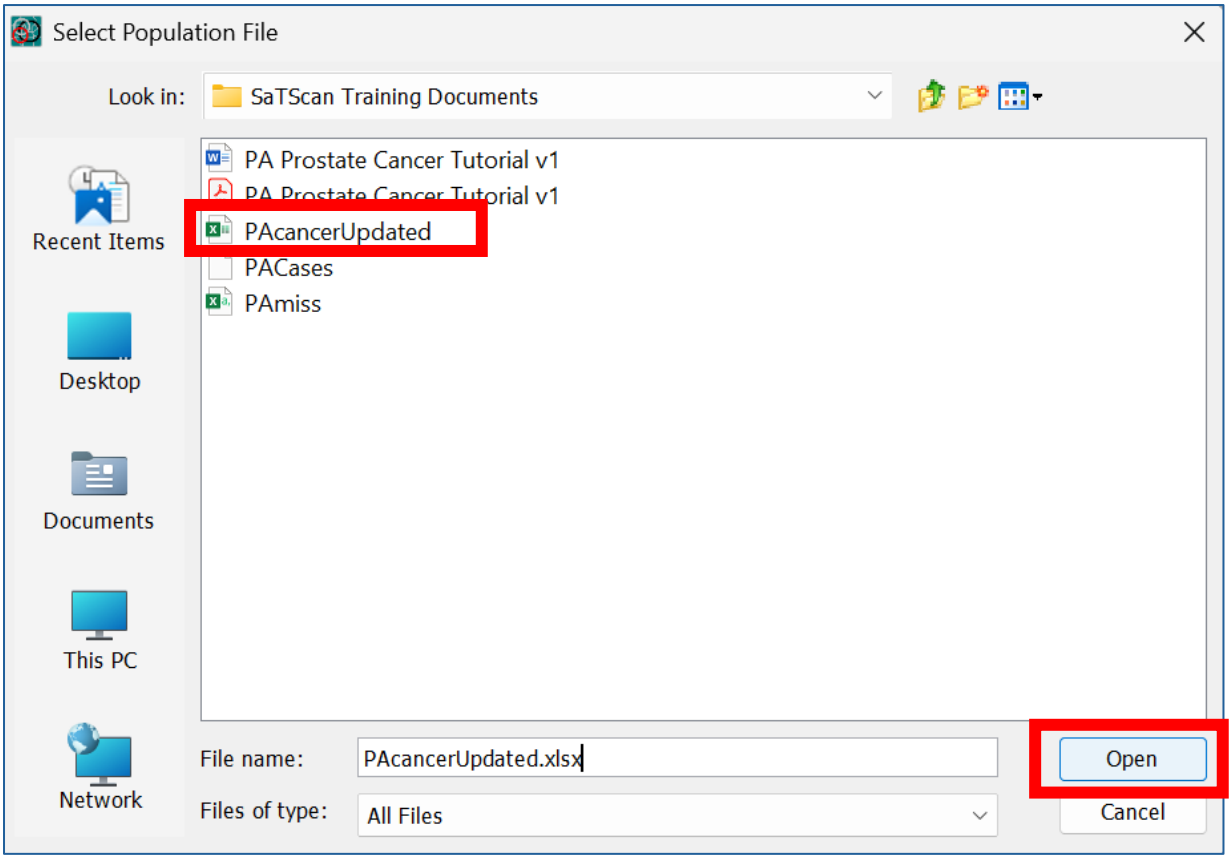


Population Data

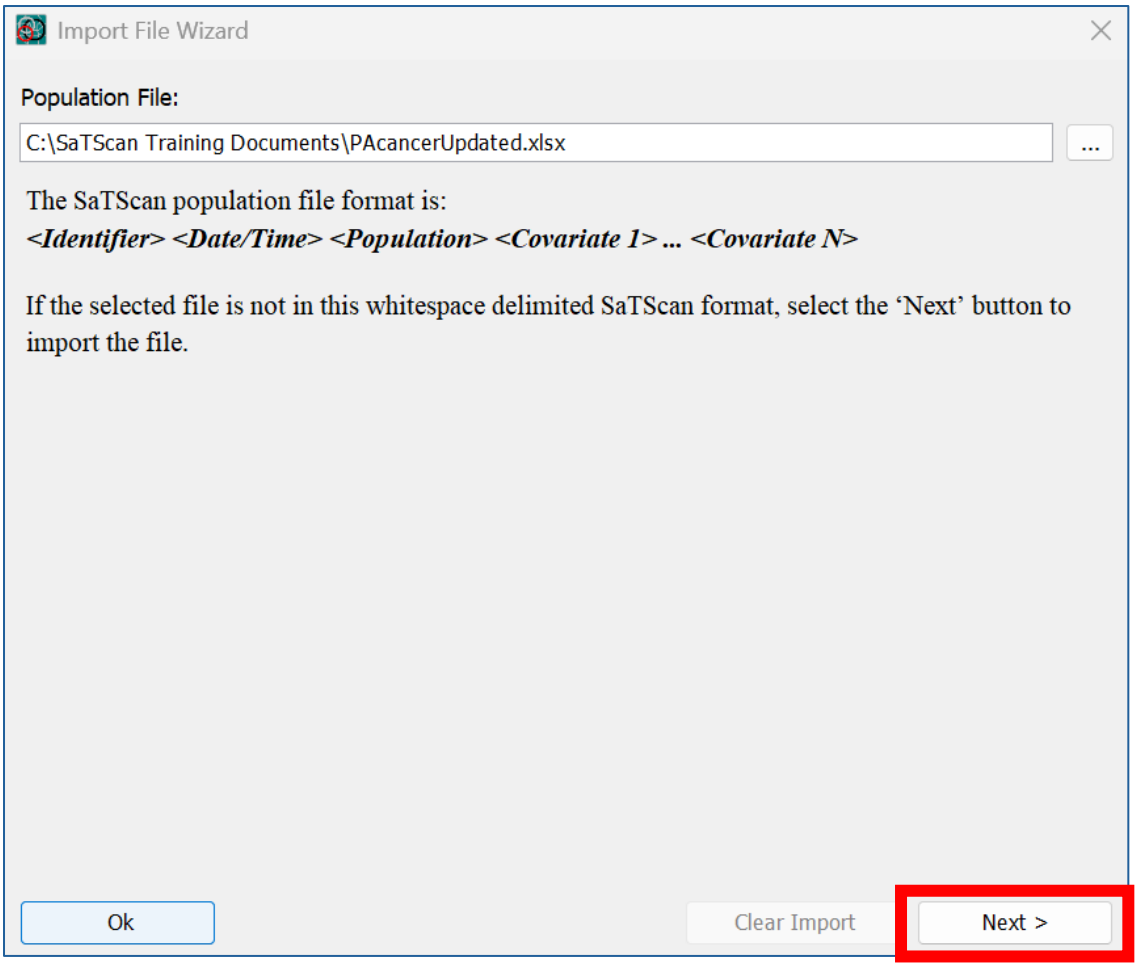
1. Next, we will follow the same process to add the **Population File**. Start by selecting the "**ellipses (…)**" next to the **Population File** line on the **Input** tab.



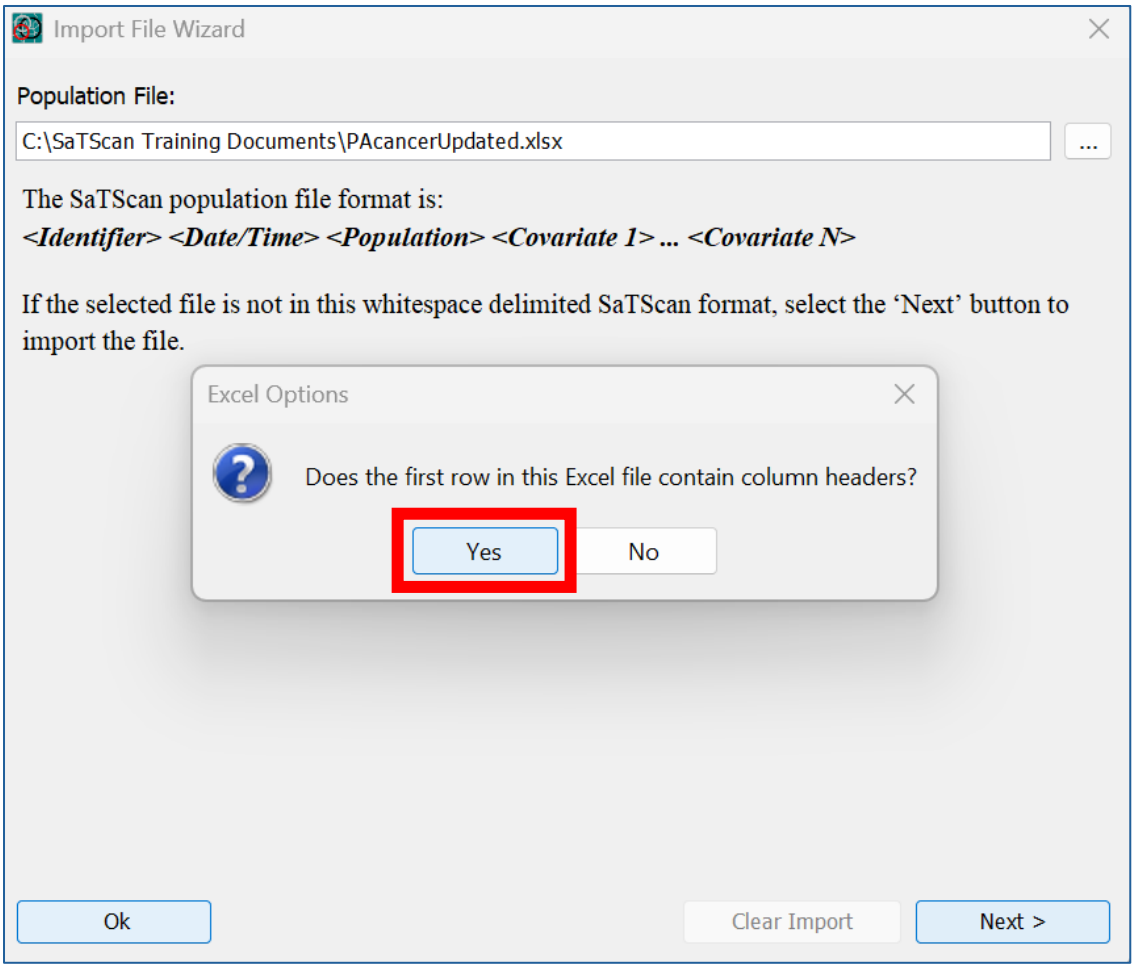
2. Next, navigate to where your tutorial data folder is stored and select the same "**PAcancerUpdated.xlsx"** file we used before for the case file. Select the "**Open**" button (in the bottom right) to select it as your **Population File**.



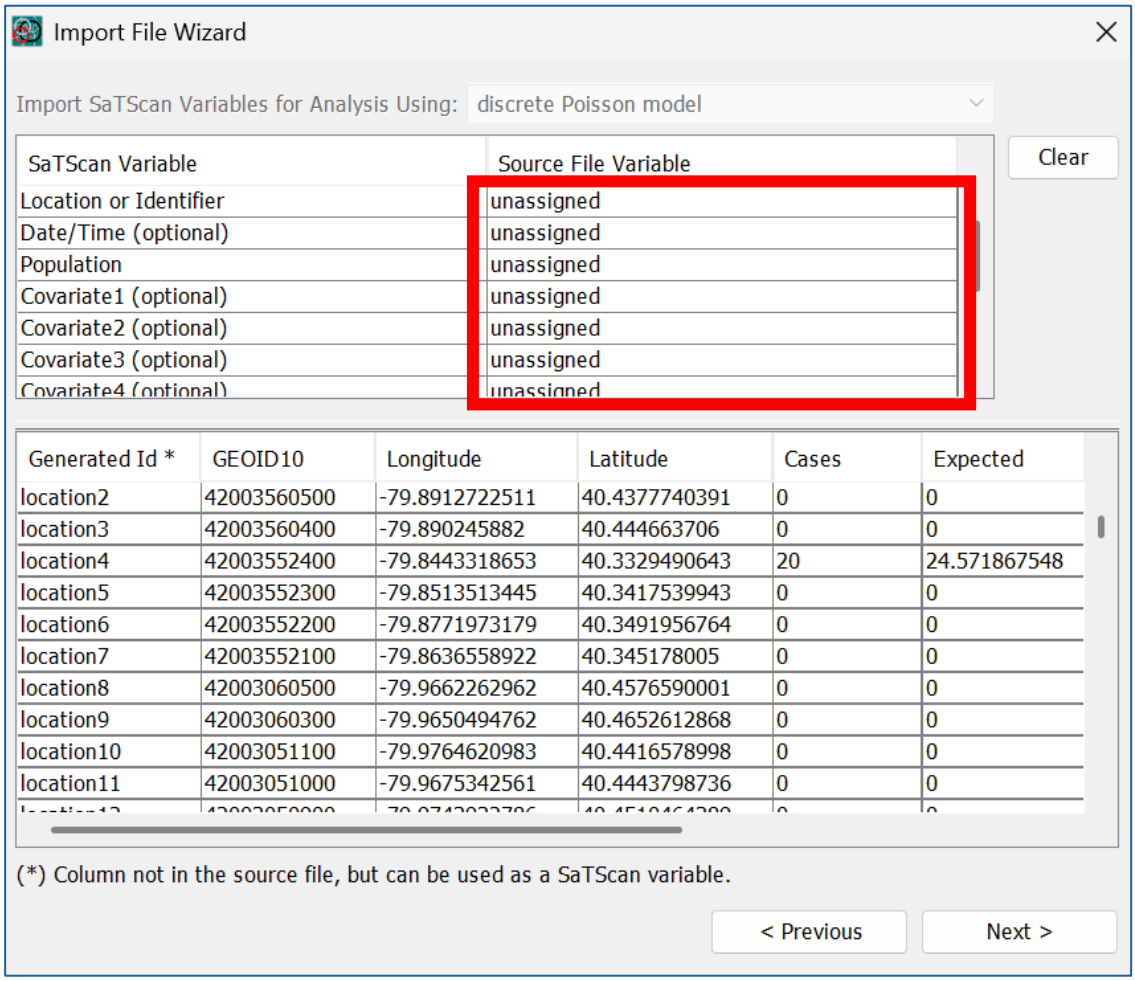
3. The next window below will appear to let you know what the file format should look like. Select the "**Next**" button to continue.



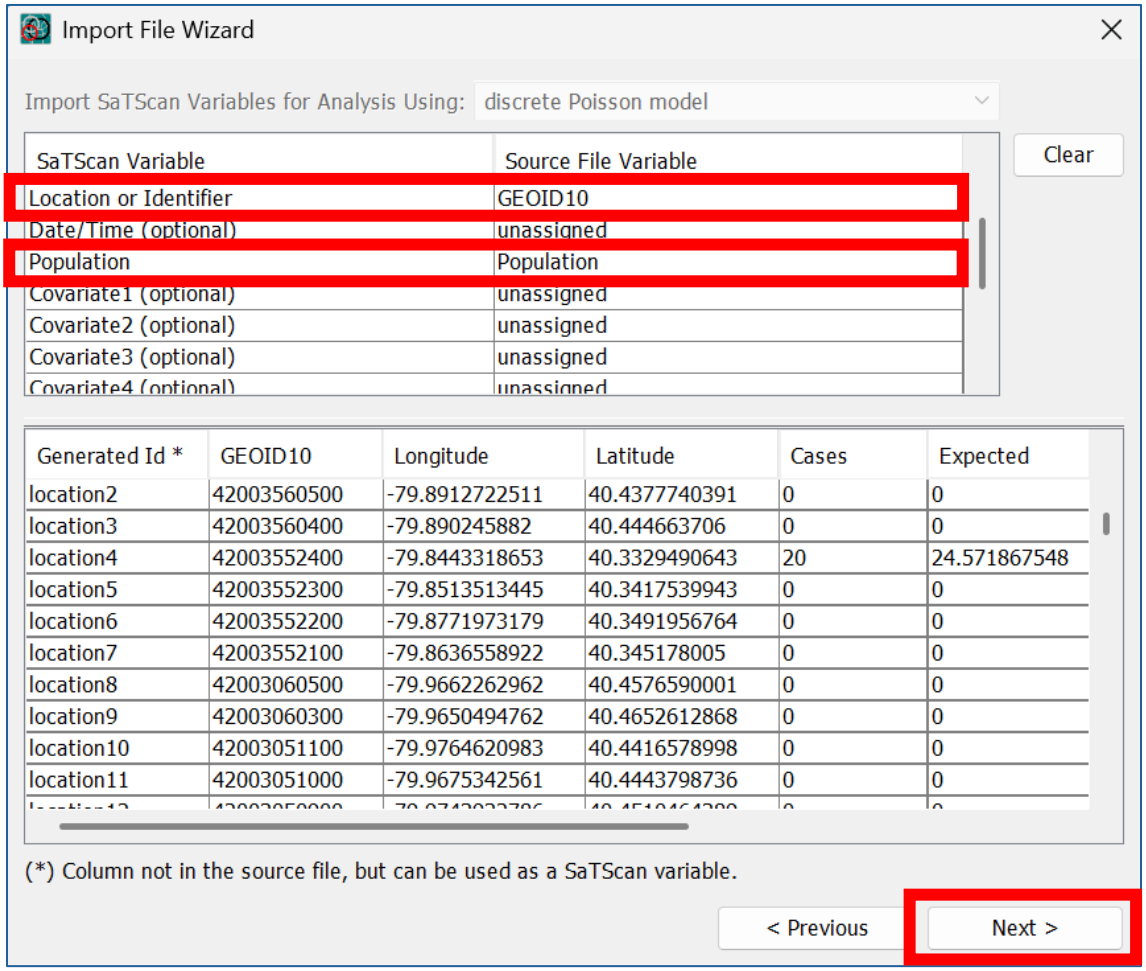
4. For Excel files (like this one) you will then get a pop-up window that asks if your file has headers. Select "**Yes**" as this file does have headers.



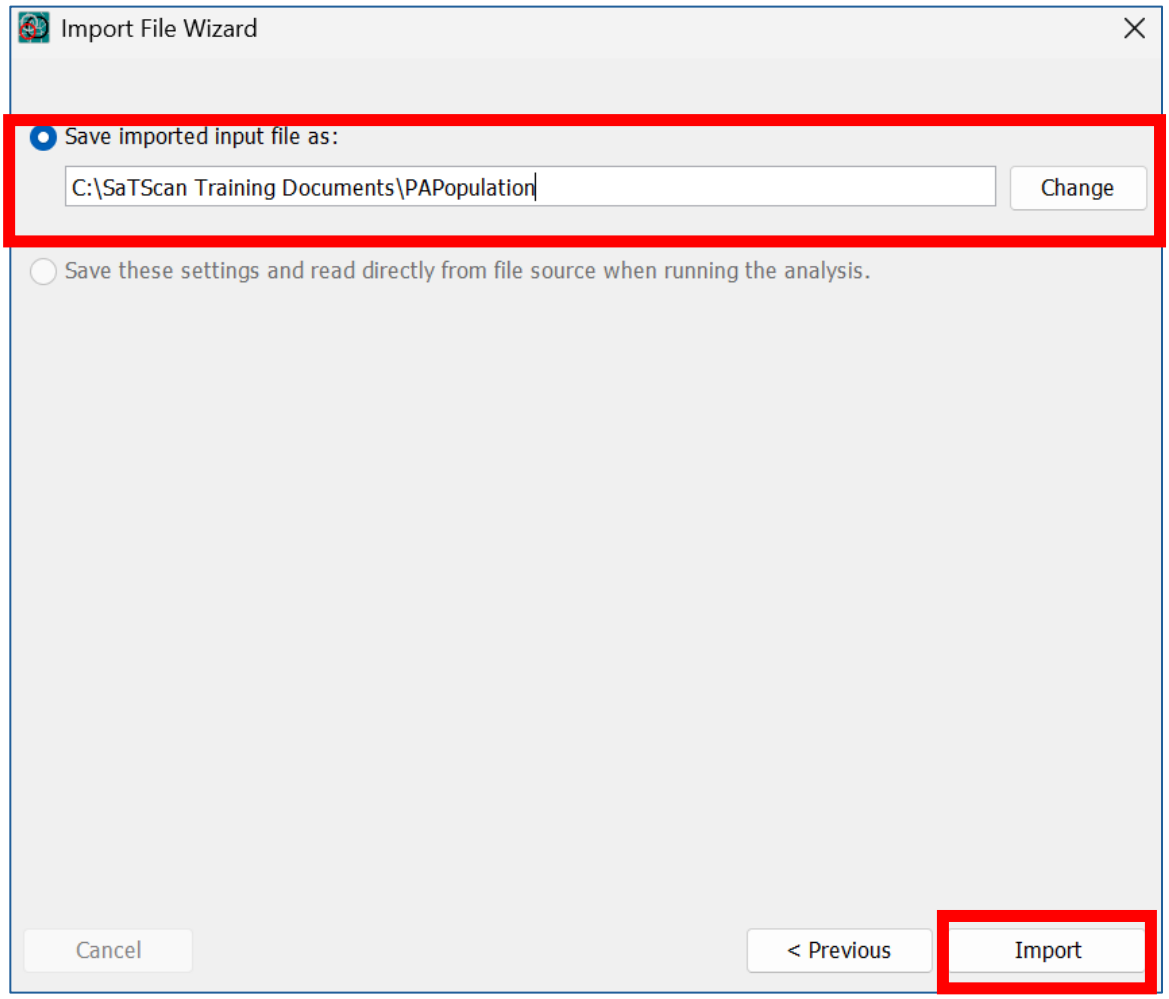
5. Next, assign each required **SaTScan Variable** we will be using for this analysis to a **Source File Variable** by clicking on the "**unassigned**" box next to each required **SaTScan Variable** and selecting the appropriate variable from the drop-down menu.



6. For this analysis, assign **[Location or Identifier]** to **[GEOID10]** and **[Population]** to the **[Population]** variable in the source file. Leave the remaining optional SaTScan variables as unassigned. Select "**Next**" to move to the next screen.

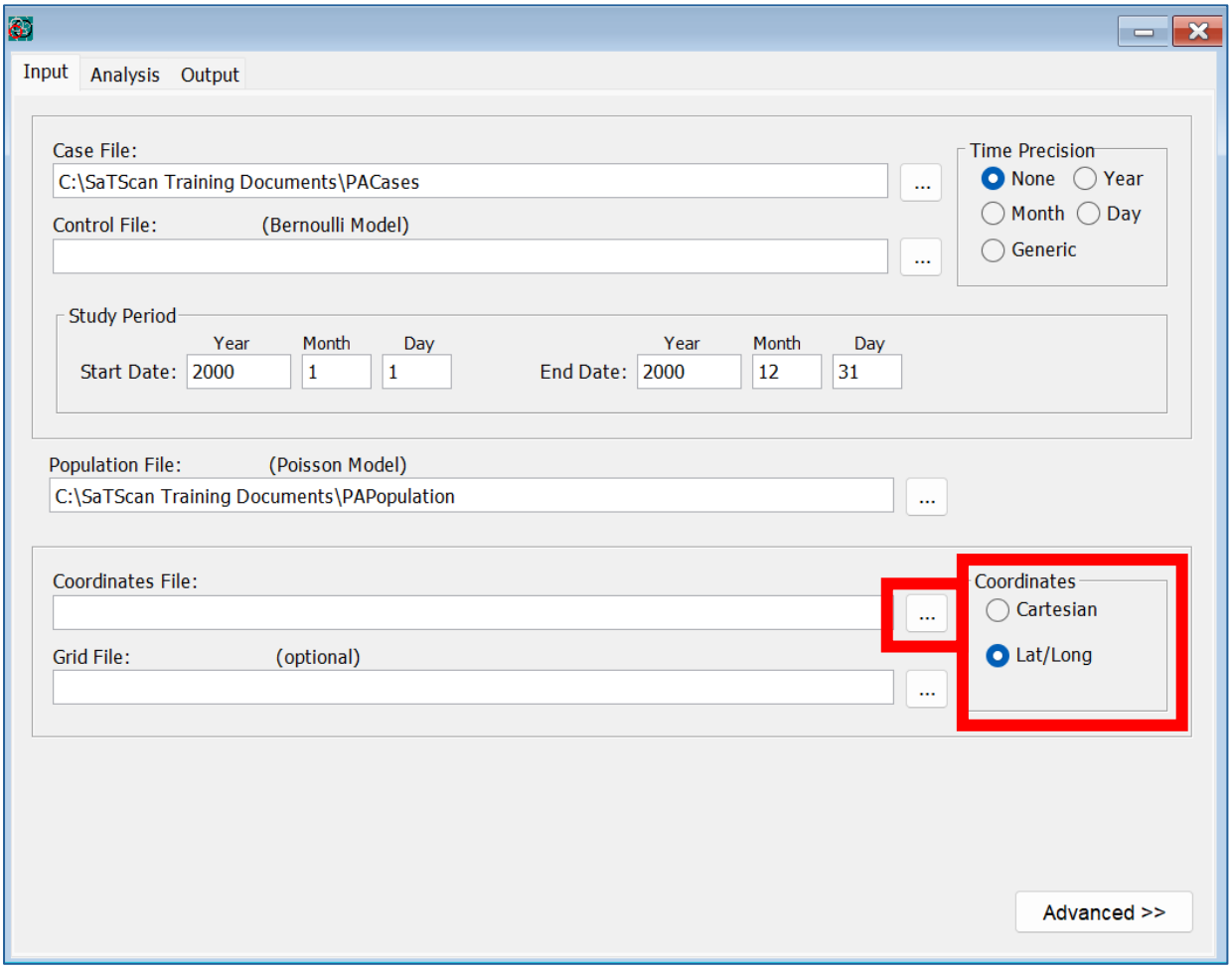


7. Next, import the variables you assigned (i.e., **[GEOID10]** and **[Population]**) into a new file that SaTScan will read the population information from. Update the file location and name in the address bar or by selecting the "**Change**" button to choose a new location. Update the file name in the address bar to "**PAPopulation**" for this tutorial. Select the "**Import**" button in the bottom right when finished.

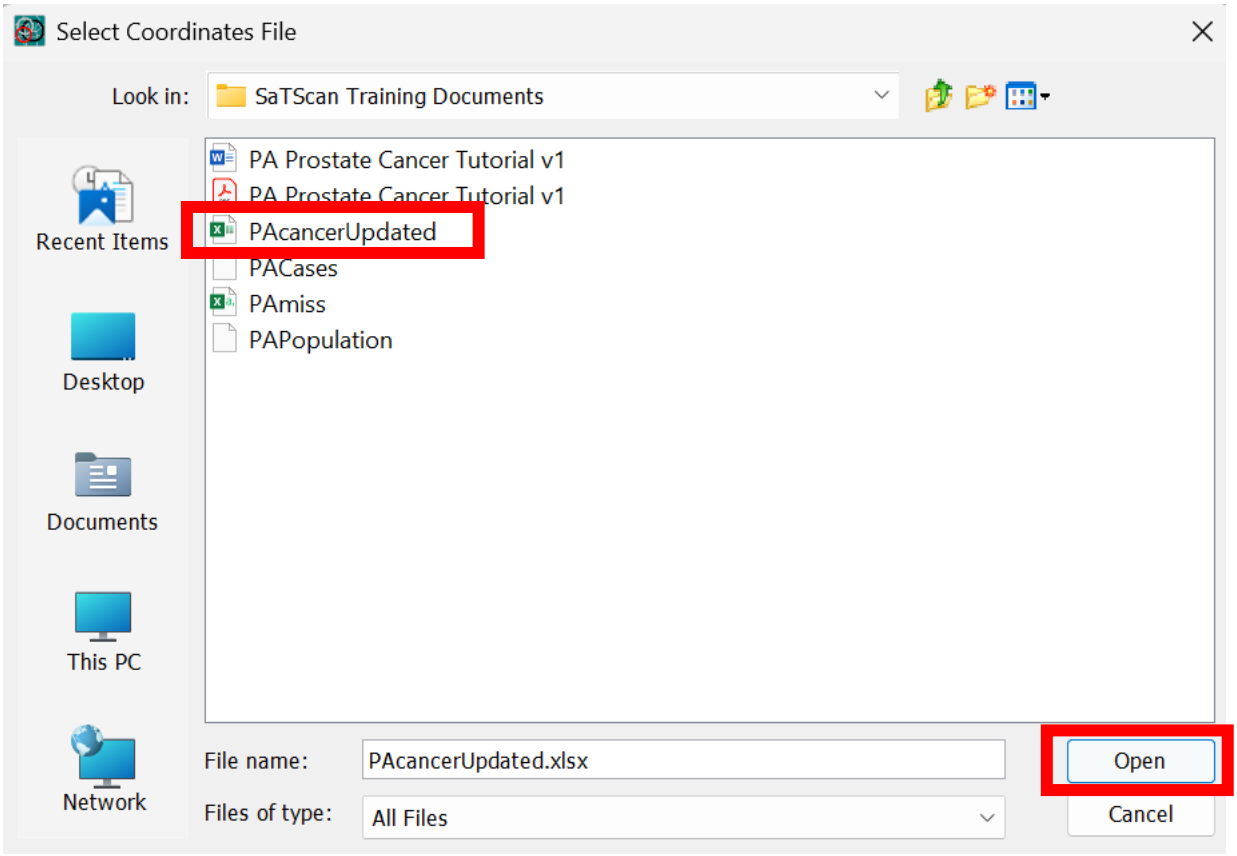


Coordinates Data

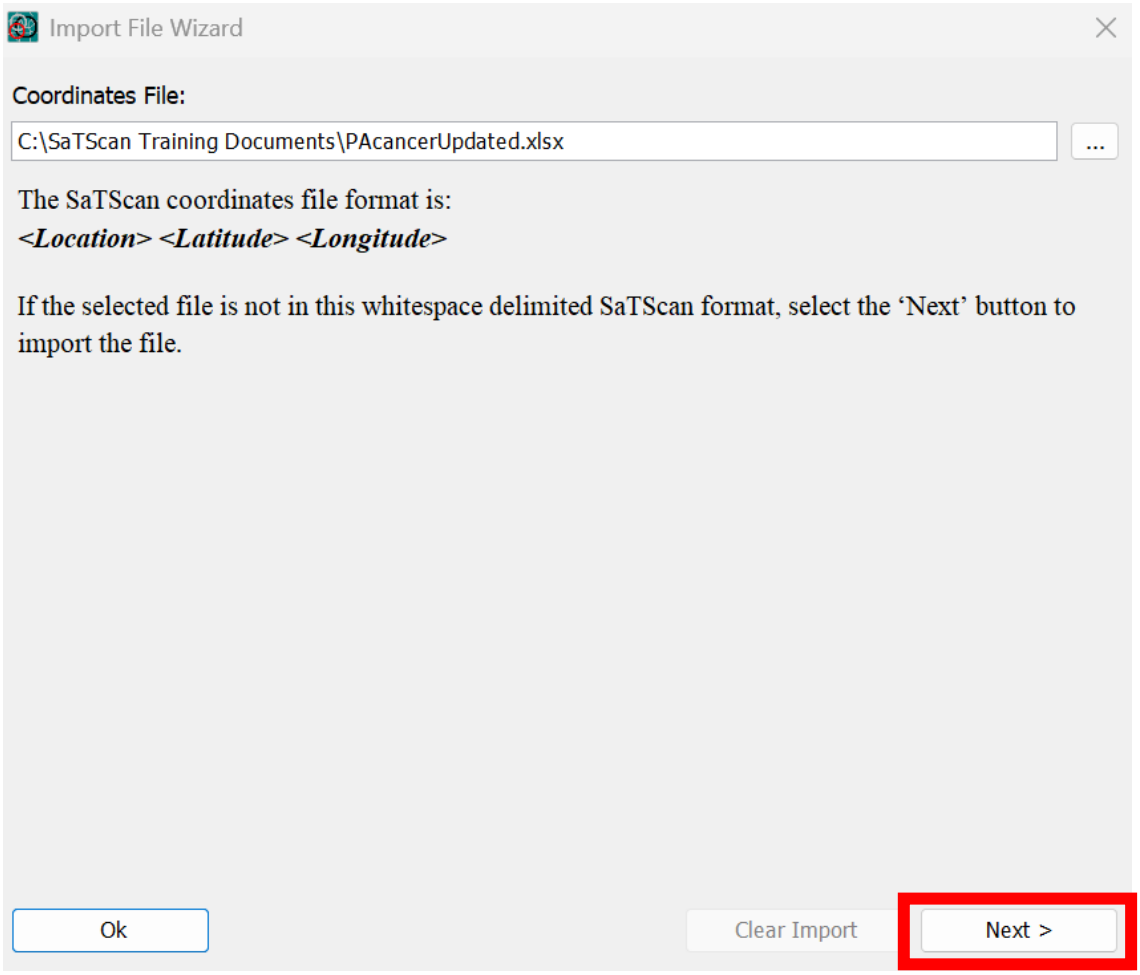
1. Lastly, we will use a similar procedure to add the **Coordinates File**. For this tutorial, the coordinates we will be using are Pennsylvania census tract centroids. On the **Input** tab, first select the coordinates system of your data. For this tutorial, choose the "**Lat/Long"** coordinates option, which may be selected by default. Next, select the "**ellipses (…)"** next to the **Coordinates File** line to begin adding your coordinates data.



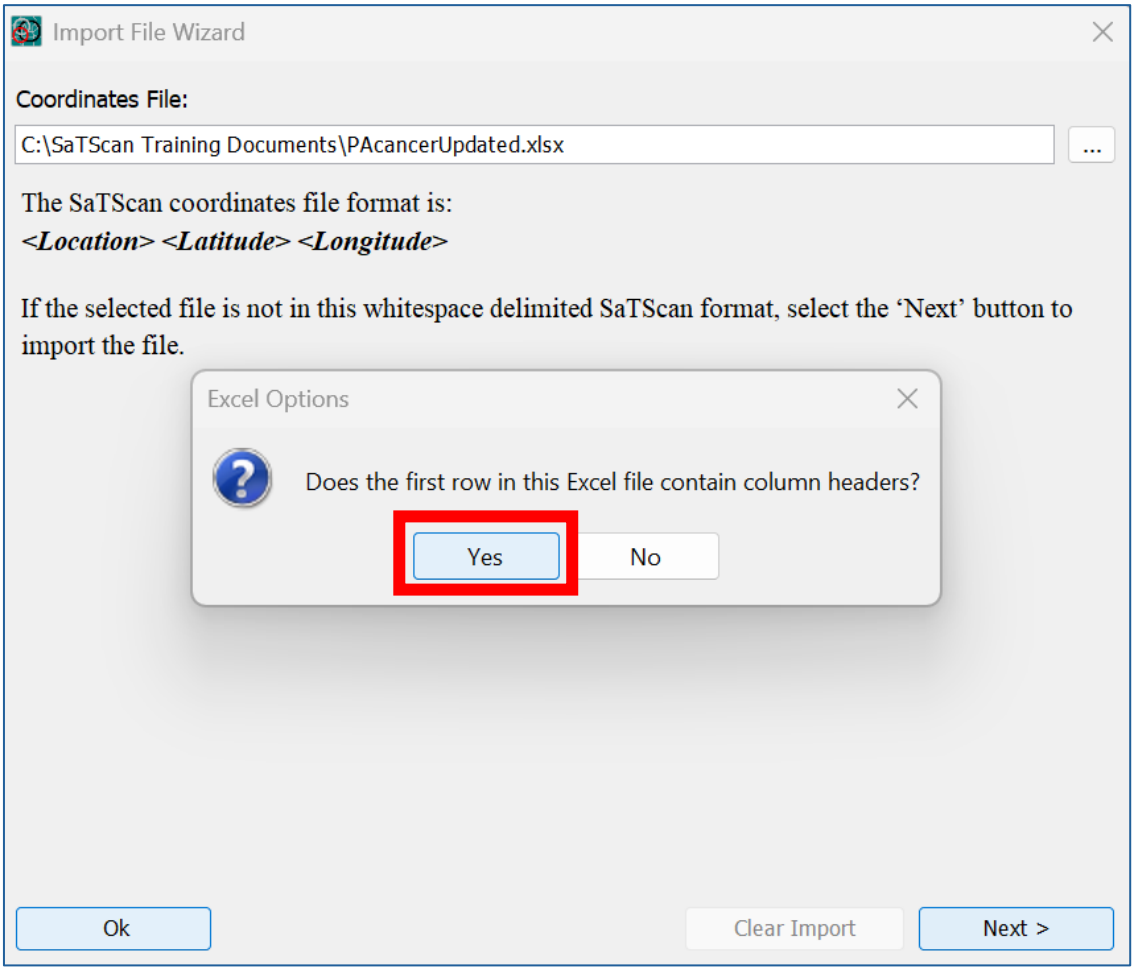
2. Next, navigate to where your tutorial data folder is stored, select the same "**PAcancerUpdated.xlsx"** file we used for the case and population files. Select the "**Open**" button (in the bottom right) to select it as your **Coordinates File**.



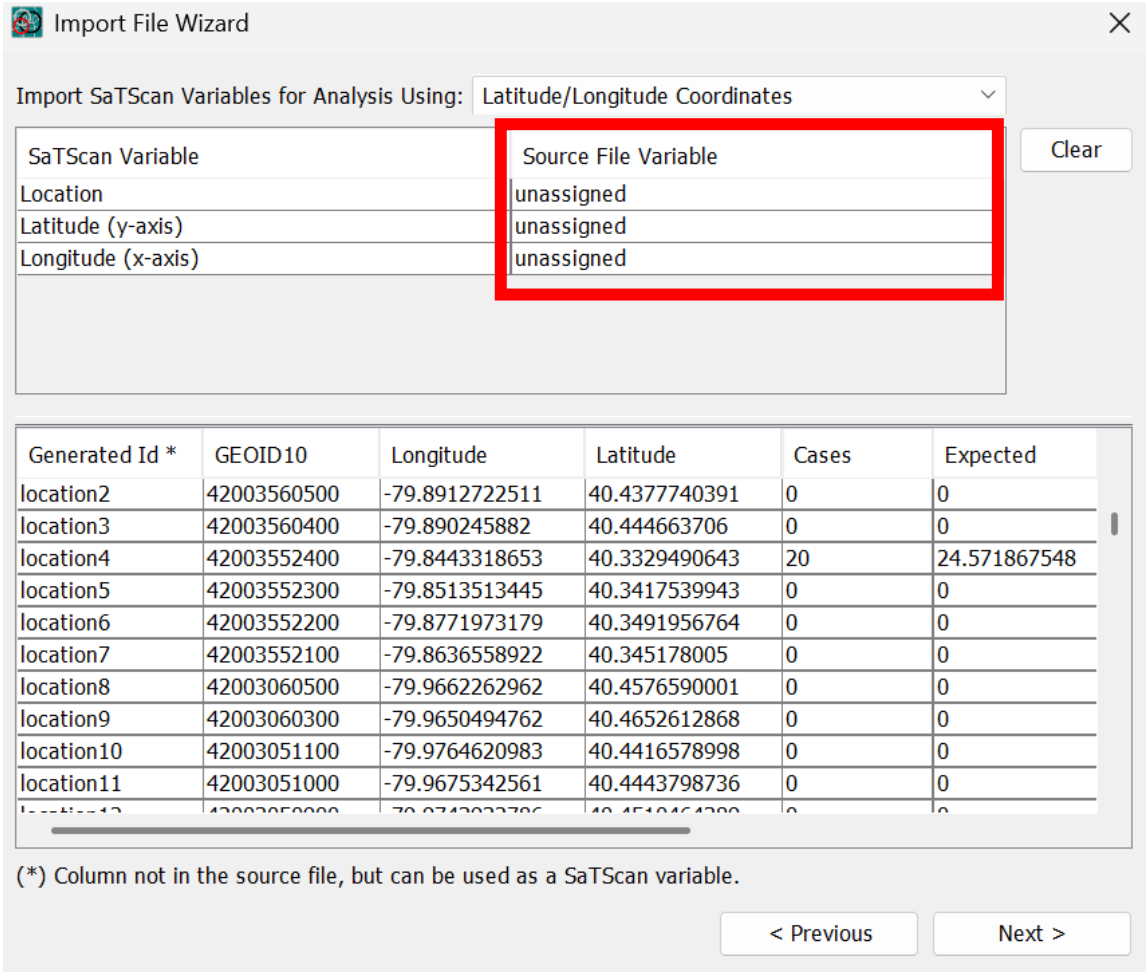
3. The next window below will appear to let you know what the file format should look like. Select the "**Next**" button to continue.



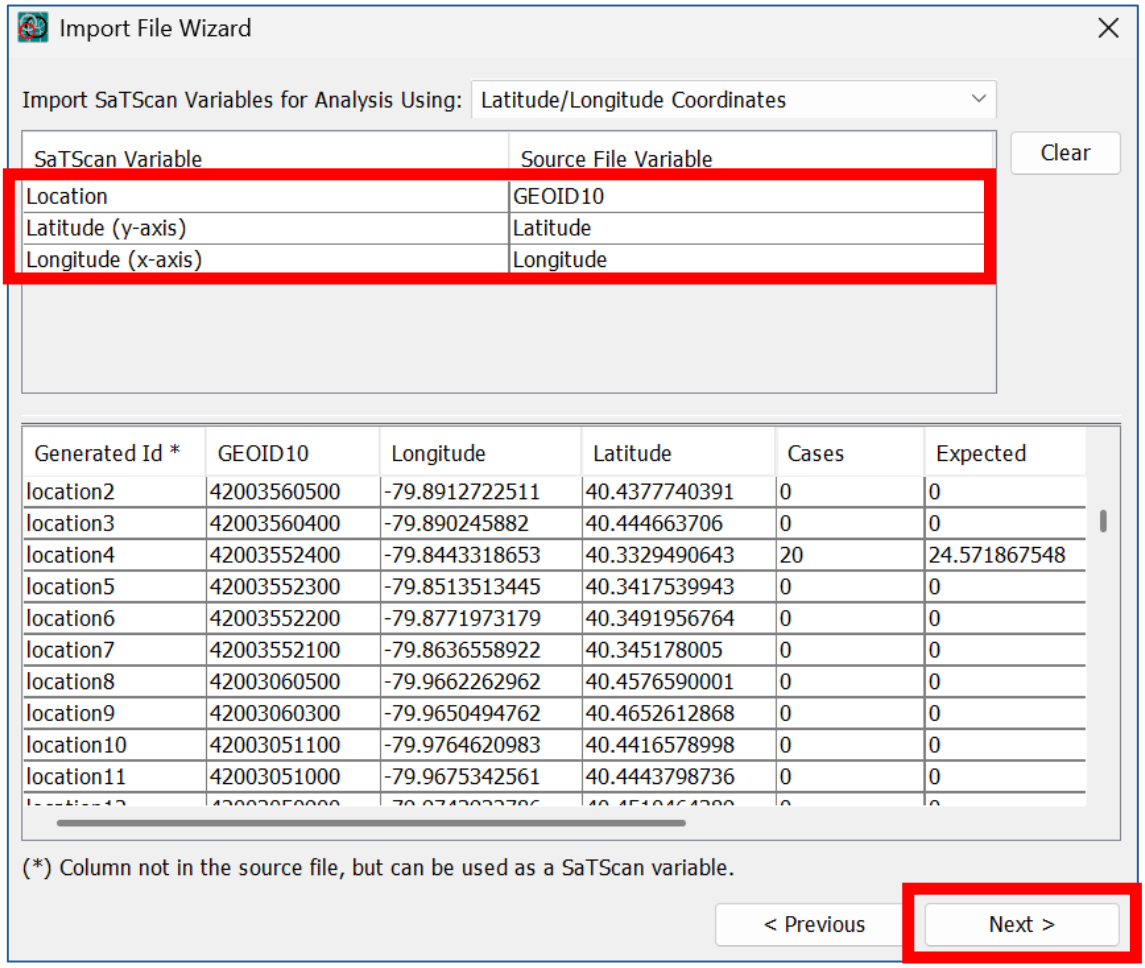
4. For Excel files (like this one) you will then get a pop-up window that asks if your file has headers. Select "**Yes**" as this file does have headers.



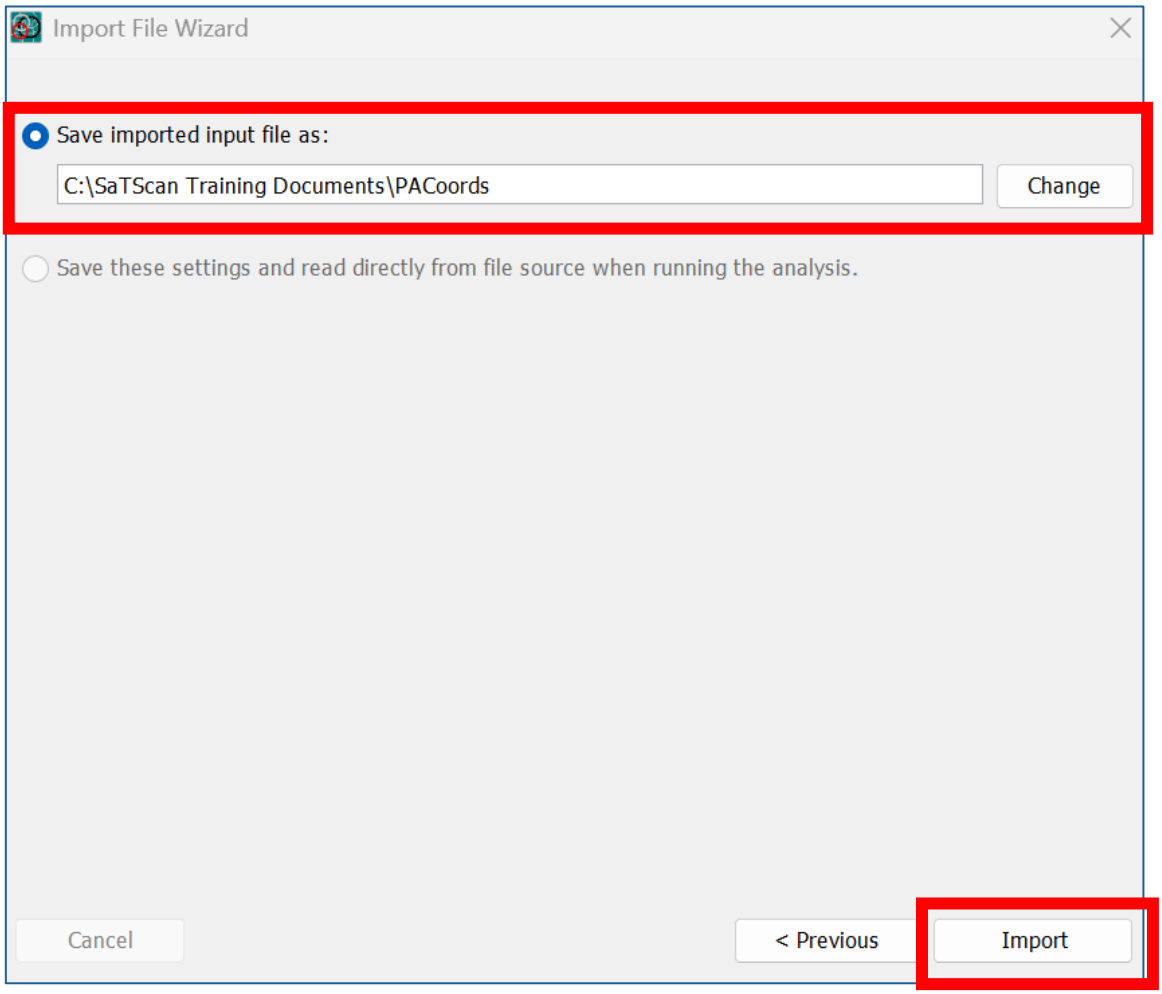
5. Next, assign each required **SaTScan Variable** we will be using for this analysis to a **Source File Variable** by selecting the "**unassigned**" box next to each of the SaTScan variables and choosing the appropriate variable from the drop-down menu.



6. Each **SaTScan Variable** is required for the **Coordinates File**, so we will need to assign a **Source File Variable** to each of them before we can move forward. Assign **[Location ID]** to **[GEOID10]**, **[Latitude (y-axis)]** to **[Latitude]**, and **[Longitude (x-axis)]** to **[Longitude]**. Select "**Next**" to move to the next screen.



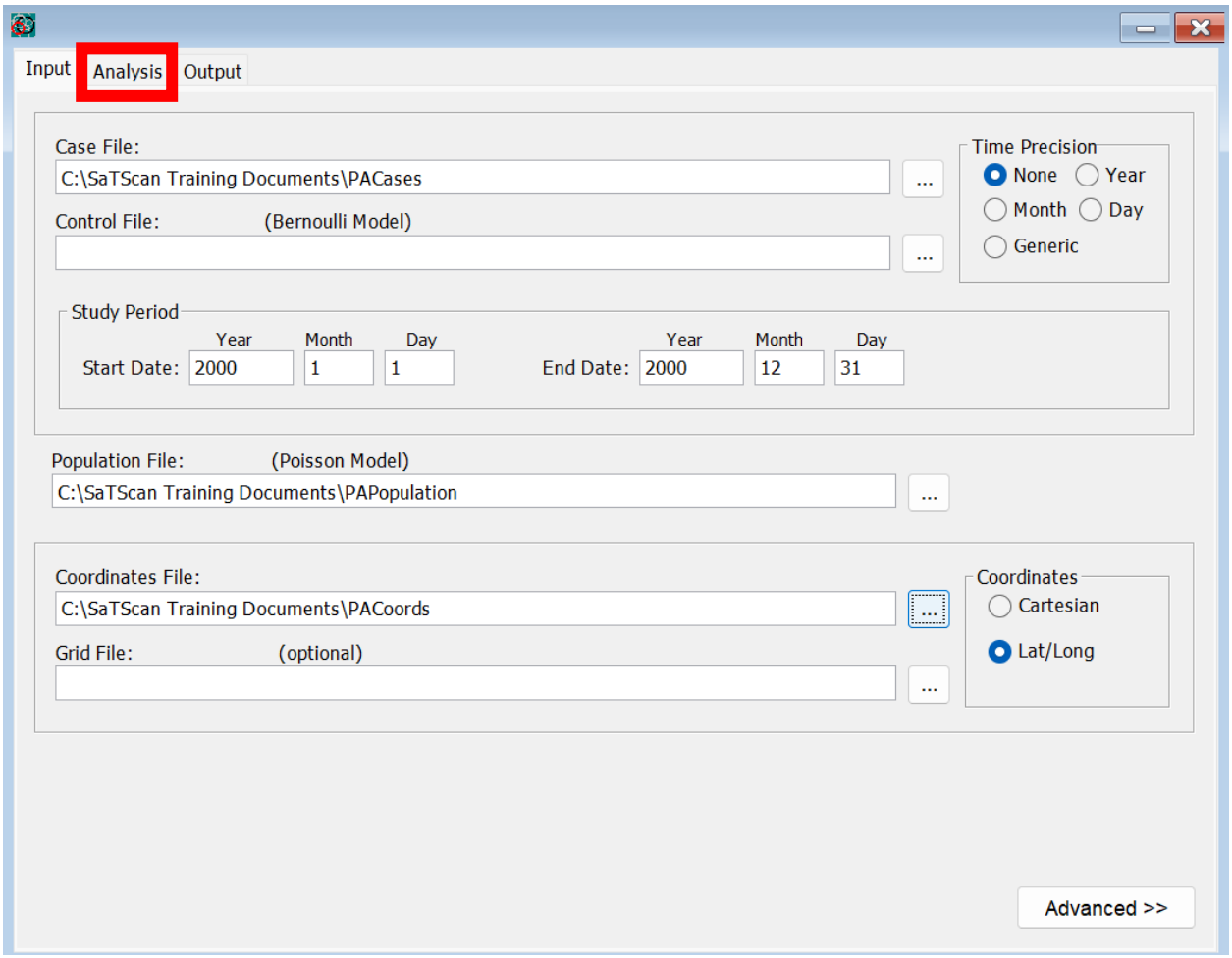
7. Import the coordinate variables you assigned into a new file that SaTScan will read the coordinates data from. Update the file location and name in the address bar or by selecting the "**Change**" button to choose a new location. Update the file name in the address bar "**PACoords**" for this tutorial. Select "**Import**" in the bottom right when finished.



Analysis

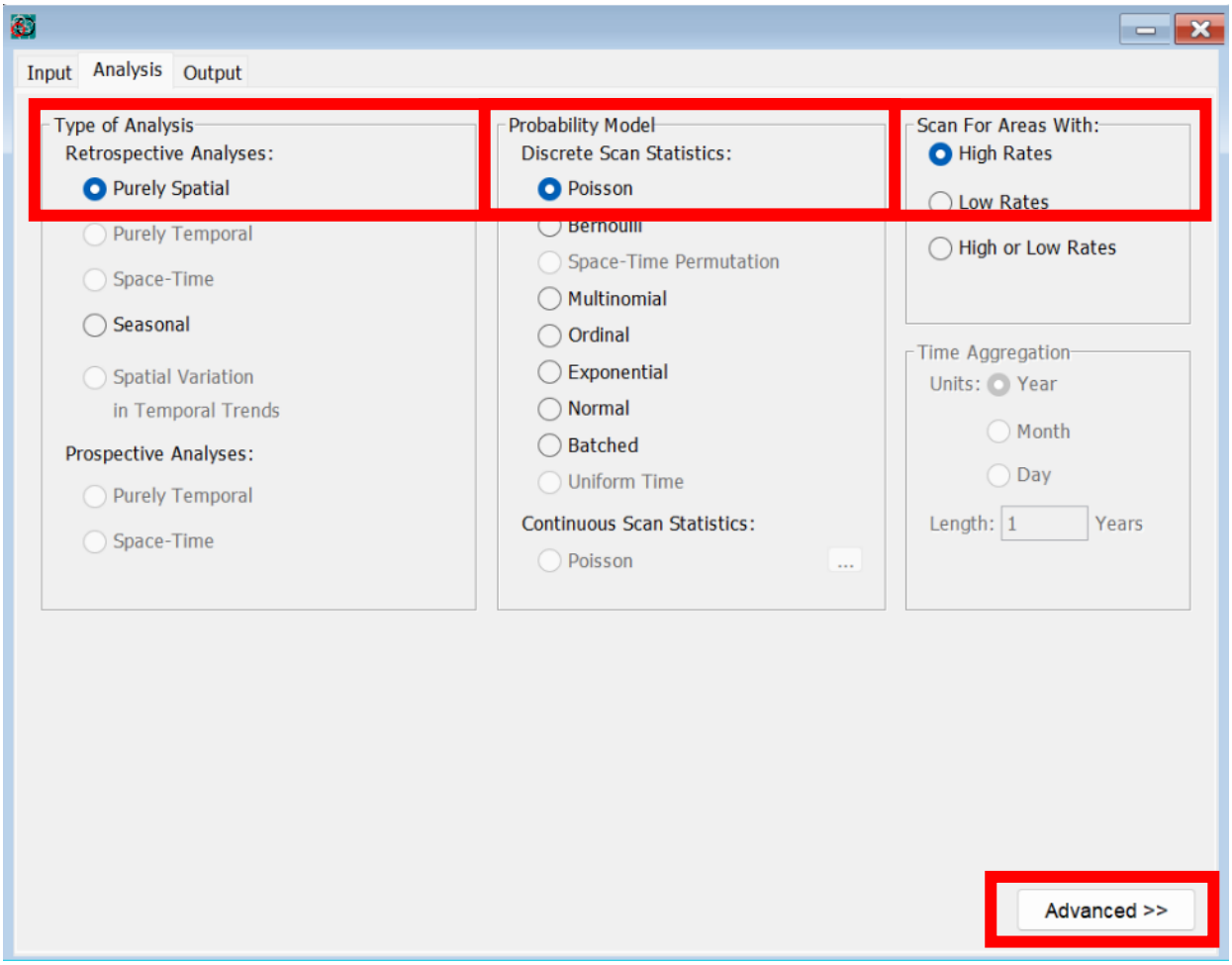
Now that we have added the required data files for our Poisson-based spatial cluster detection analysis, we can now focus on setting the parameters for our analysis. We can specify the type of analysis, the probability model that we will be using, and what clusters to search for in the **Analysis** tab.

1. Select the "**Analysis"** tab at the top of the screen.



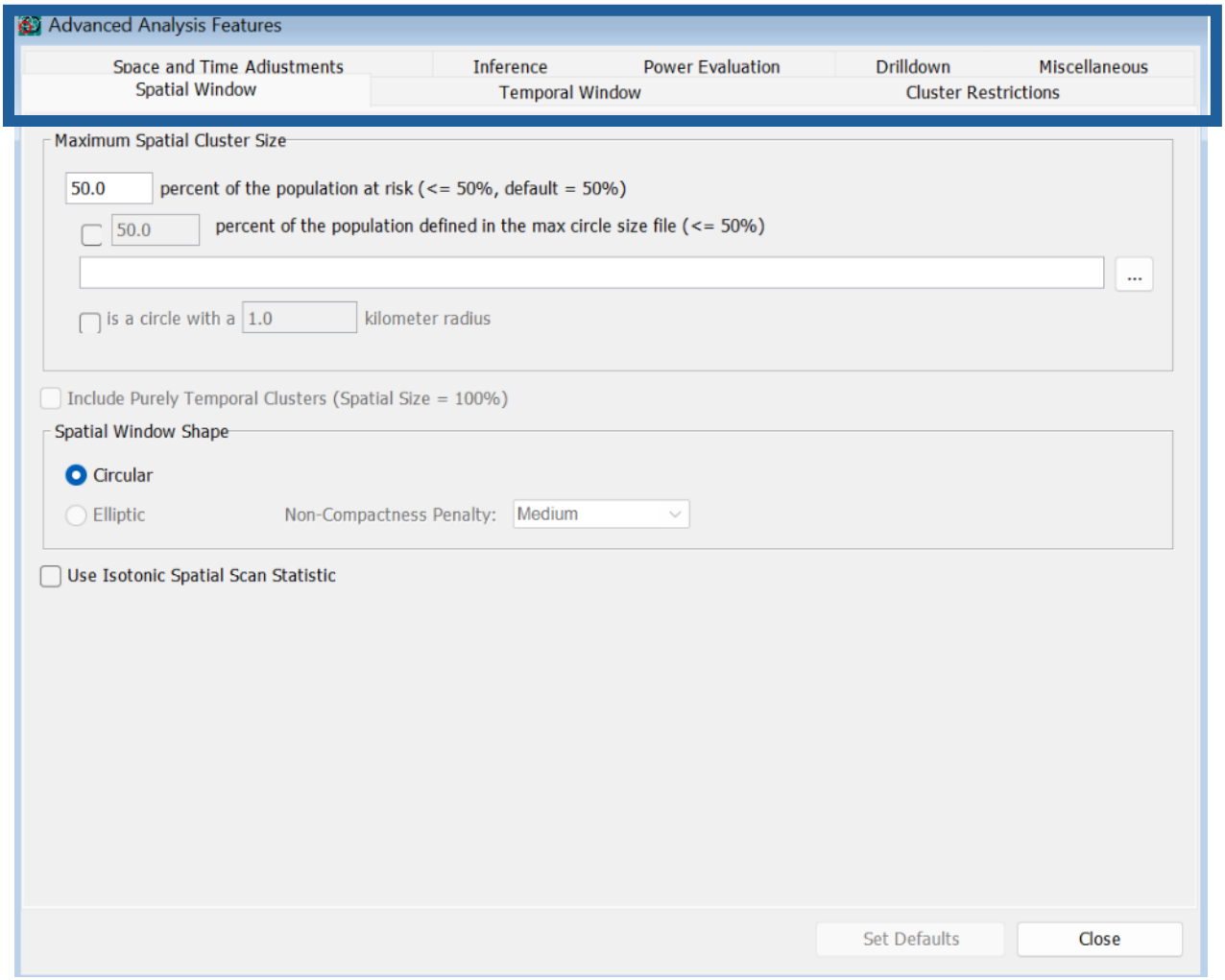
2. Next, for **Type of Analysis** select "**Purely Spatial**", for **Probability Model** select "**Poisson**", and for the **Scan for Areas With** menu select "**High Rates**". Then select the "**Advanced**" button in the bottom right of the window to set additional parameters.

   ***Note**: Some of these parameters may already be selected by default. We can also ignore the other options for this analysis, including the Time Aggregation section, as we are not performing a space-time analysis.*



Advanced Analysis – Maximum Spatial Cluster Size & Missing Data

There are several advanced features that can be set in each of the tabs at the top of the window below. For this tutorial, we will show you how to set the **maximum spatial cluster size** and how to instruct SaTScan to handle **missing data**.

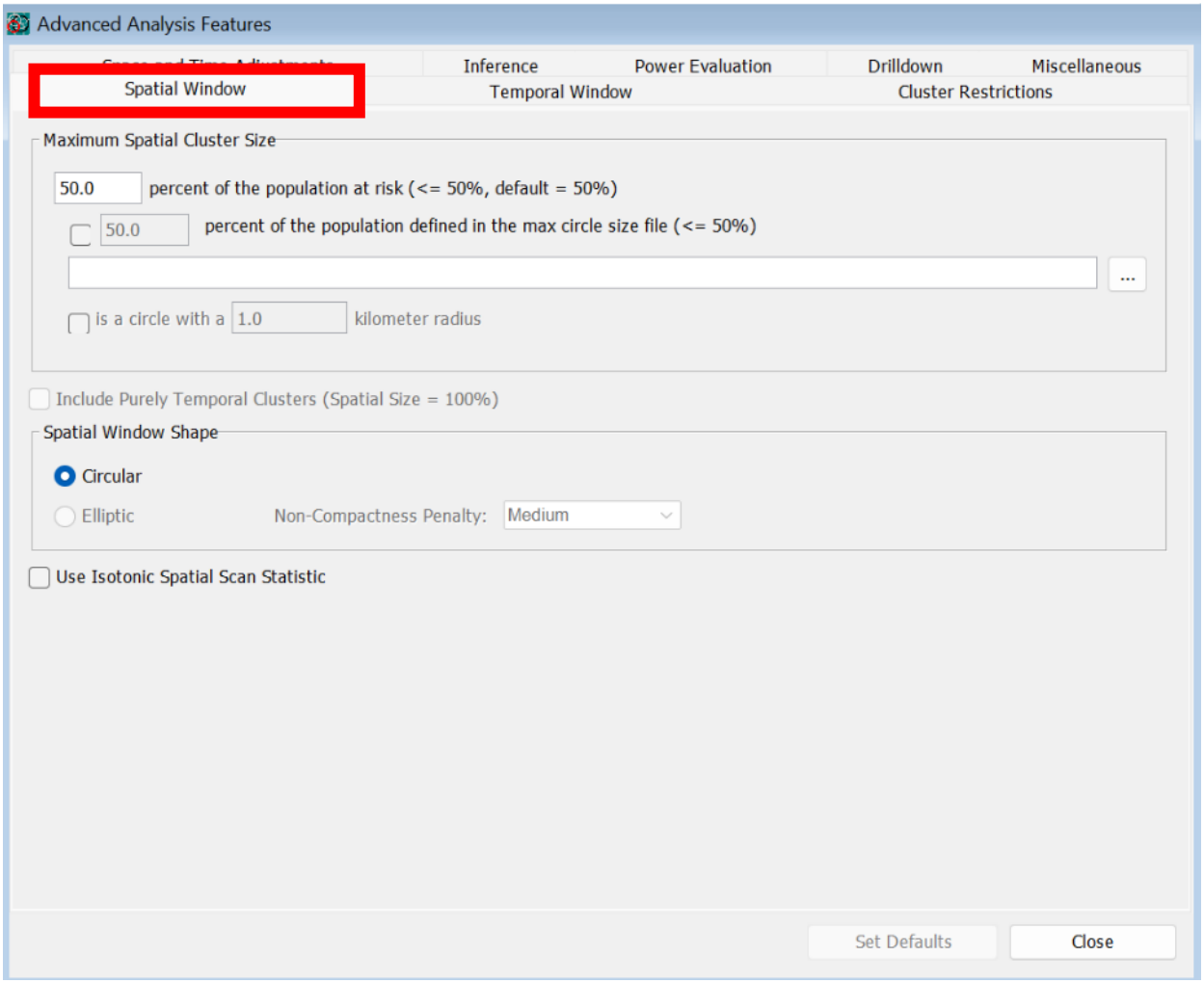


Maximum Spatial Cluster Size

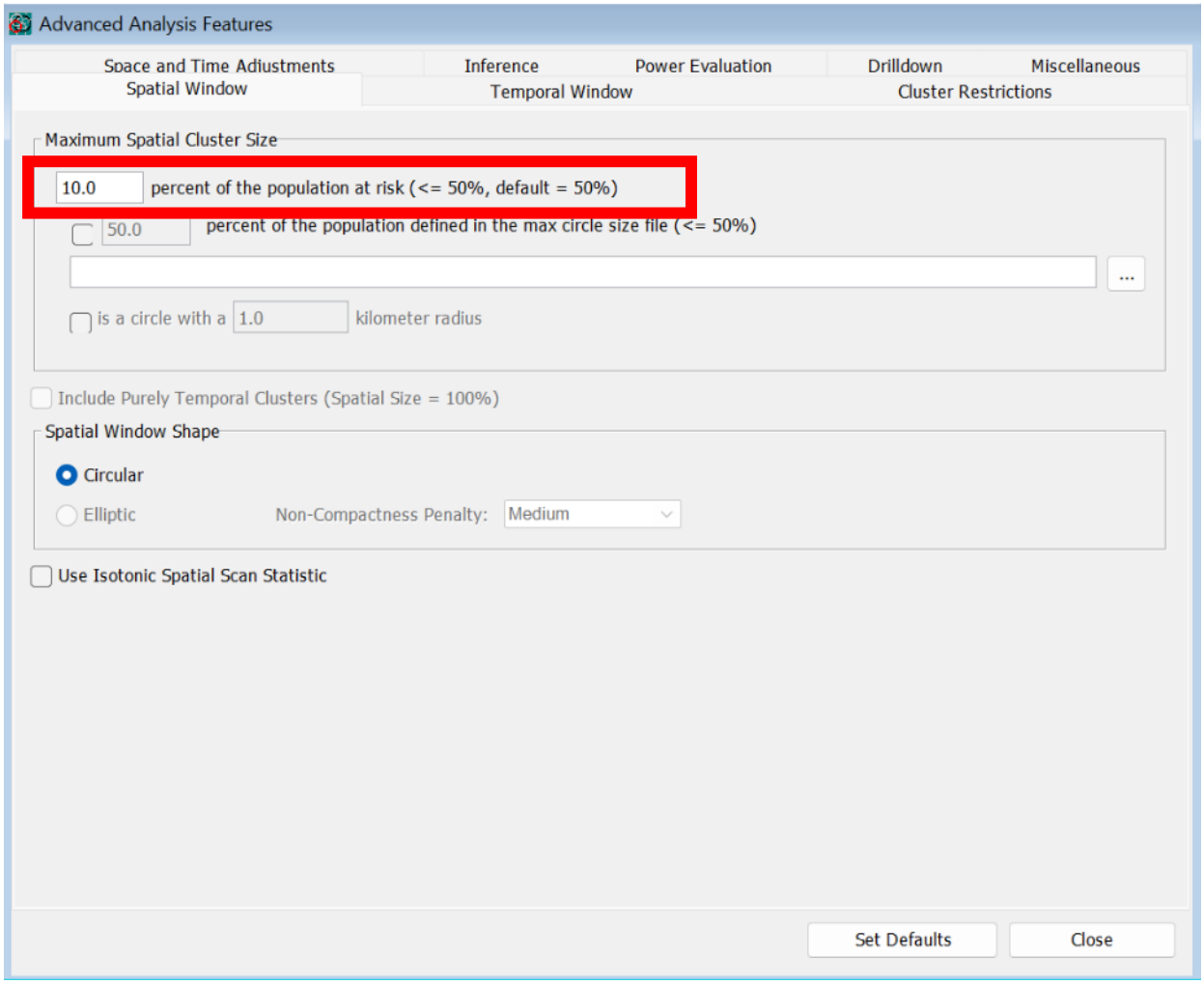
The **maximum spatial cluster size** is defined as the maximum size of any cluster that SaTScan identifies. Maximum spatial cluster size is typically defined as a percentage of the total population in the study area.

SaTScan uses a default cluster size of 50% (i.e., 50% of the total population in the study area), which is often too large of a cluster size to be meaningful for many spatial analyses. To avoid this issue and get more meaningful results, users typically set the spatial cluster size to be even smaller, such as somewhere between 10-25%. As we are focused on only a subset of the PA population today (i.e., males 18 years or older diagnosed with prostate cancer during 2010-2019), we will set the maximum spatial cluster size to 10%.

1. To set the maximum spatial cluster size, first select the "**Spatial Window**" tab at the top of the screen.



2. Adjust the default maximum cluster size from 50% to 10%, by entering "**10.0**" in the box in the **Maximum Spatial Cluster Size** section.

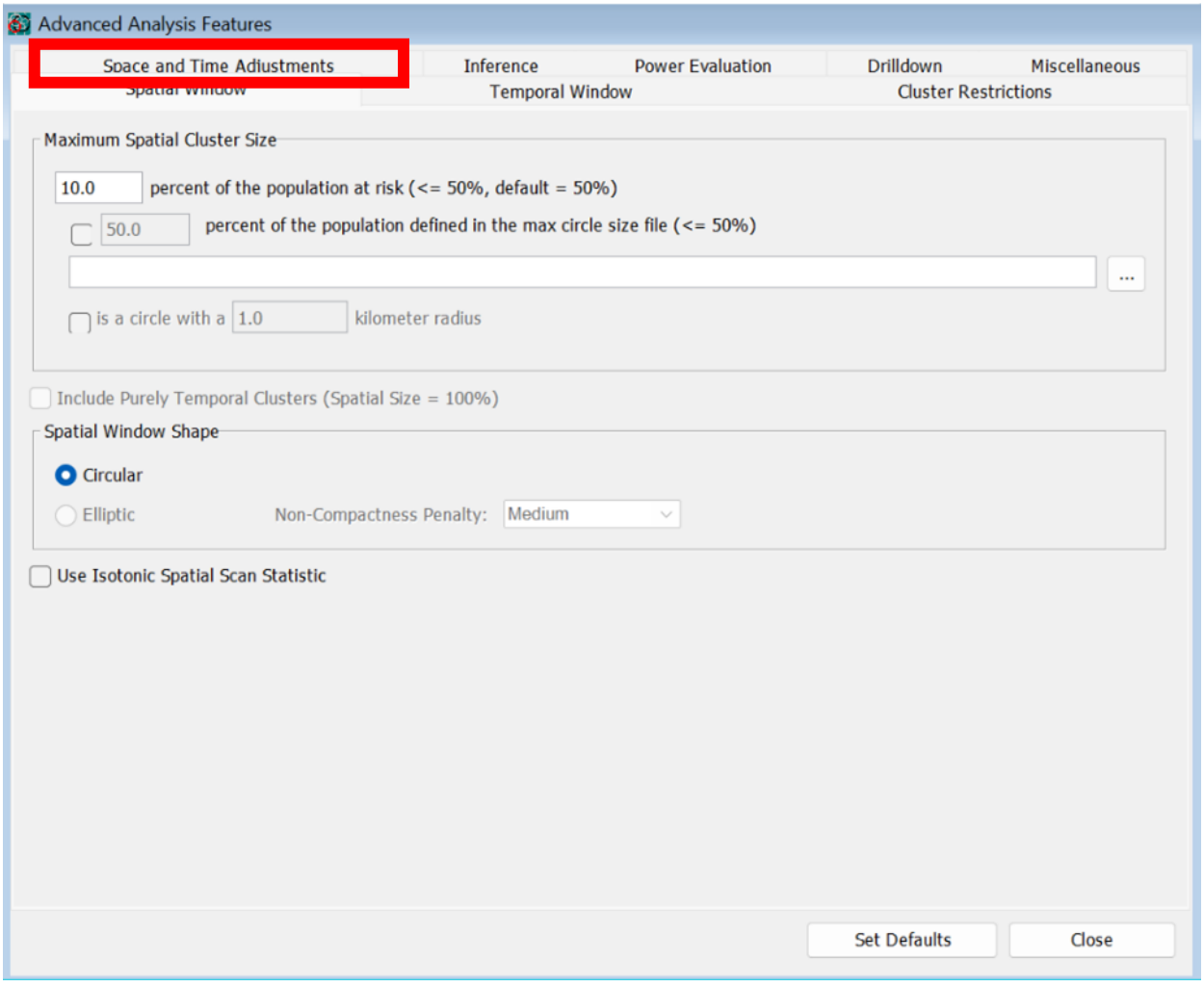


Missing Data

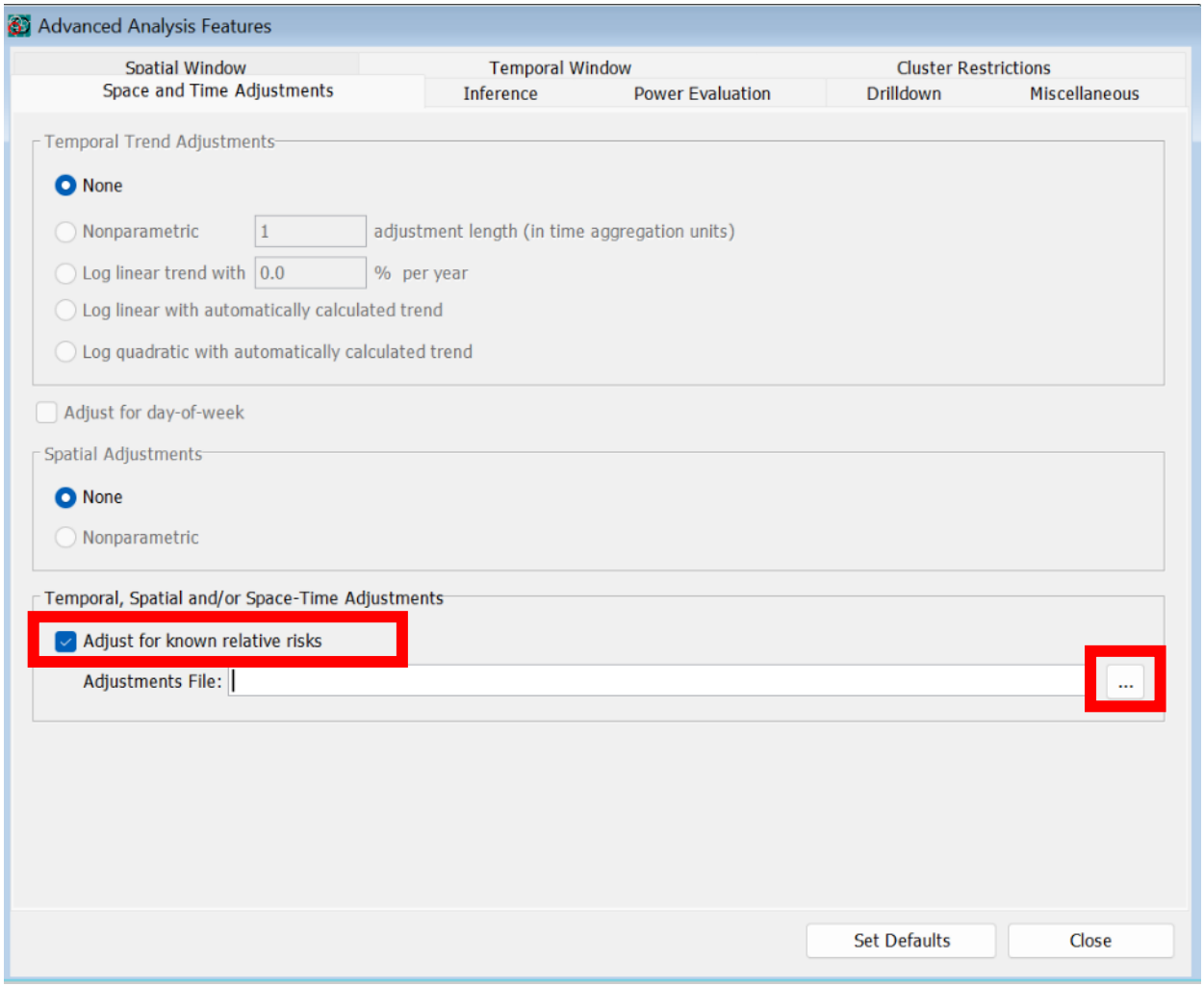
It is common for health outcome data, such as cancer rates, to be missing or suppressed due to small numbers – particularly when using smaller geographies (e.g., census tracts or block groups). SaTScan can account for missing and/or suppressed data as long as we specify how it should handle the data by creating and importing a missing data file (i.e., an "**adjustment file**").

Today's tutorial dataset is **missing data** for **701 tracts (~20% of all PA tracts**). For the Poisson model we are using today, the **adjustment file** needs to contain **2 variables:** (1) **a location variable** that corresponds to the geographic identifiers we use in the analysis files, such as GEOID10, and (2) **a relative risk variable** that contains a "0" value for all geographies with missing or suppressed data.

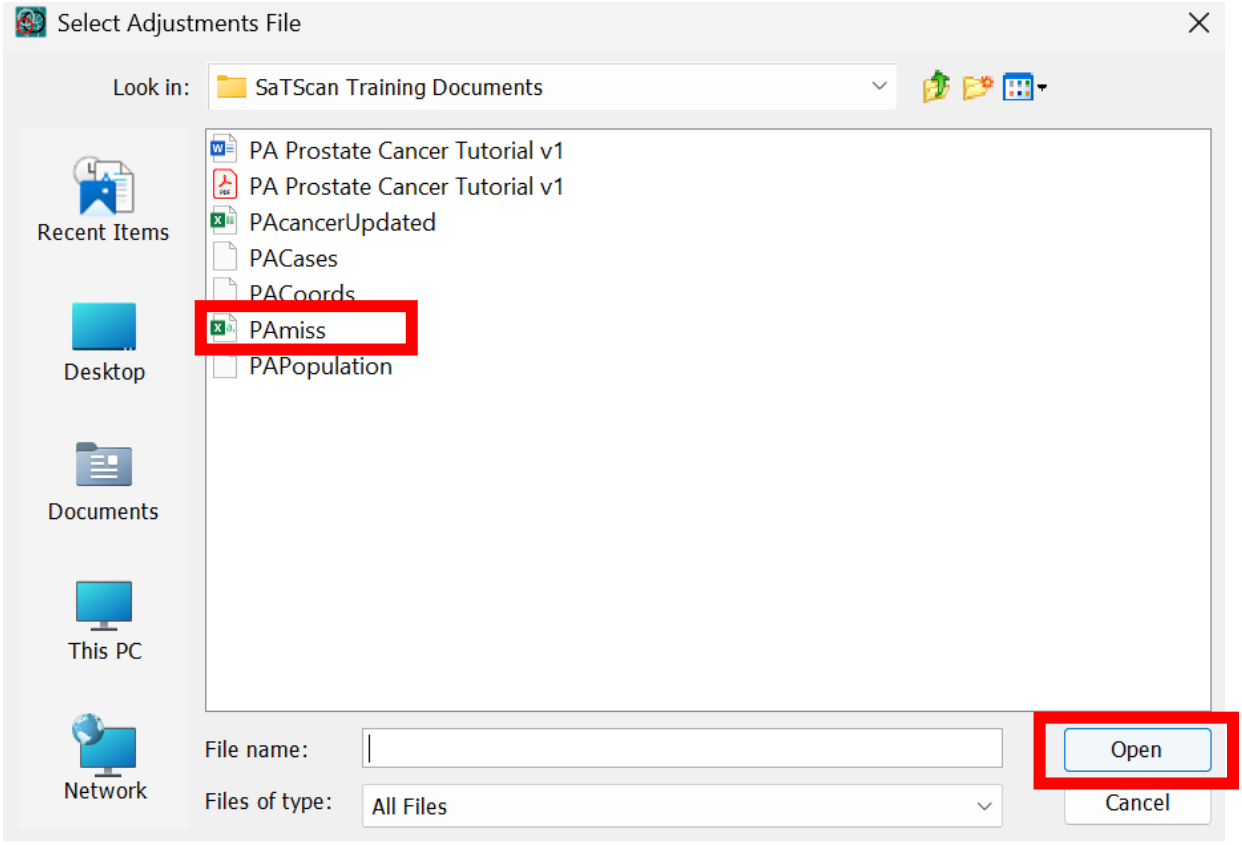
1. Tell SaTScan how to handle missing data, by first selecting the "**Space and Time Adjustments**" tab at the top of the screen.



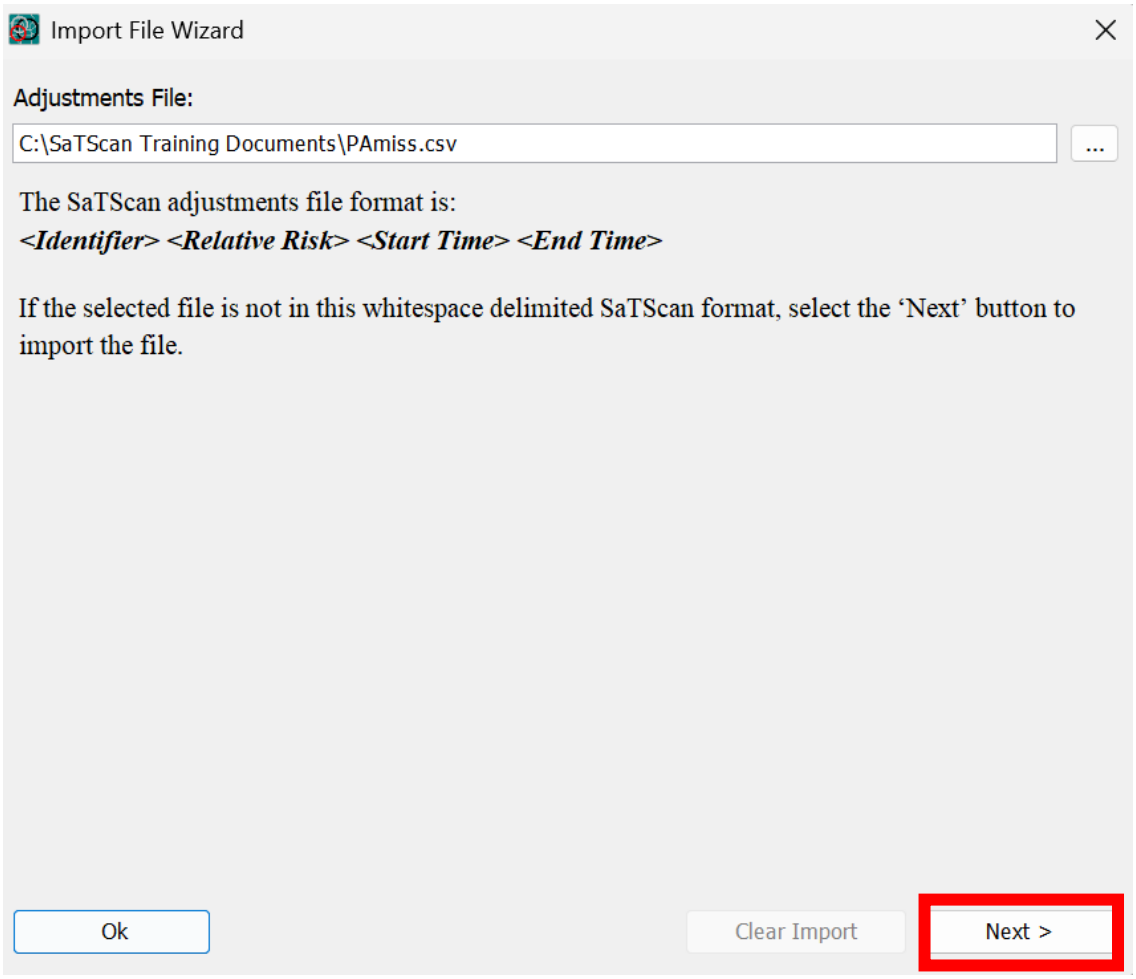
2. Next, check the box next to "**Adjust for known relative risks**" in the **Temporal, Spatial and/or Space-Time Adjustments** section. Then tell SaTScan where to find your adjustment file, by selecting the "**ellipses (…)**" next to the **Adjustment File** line.



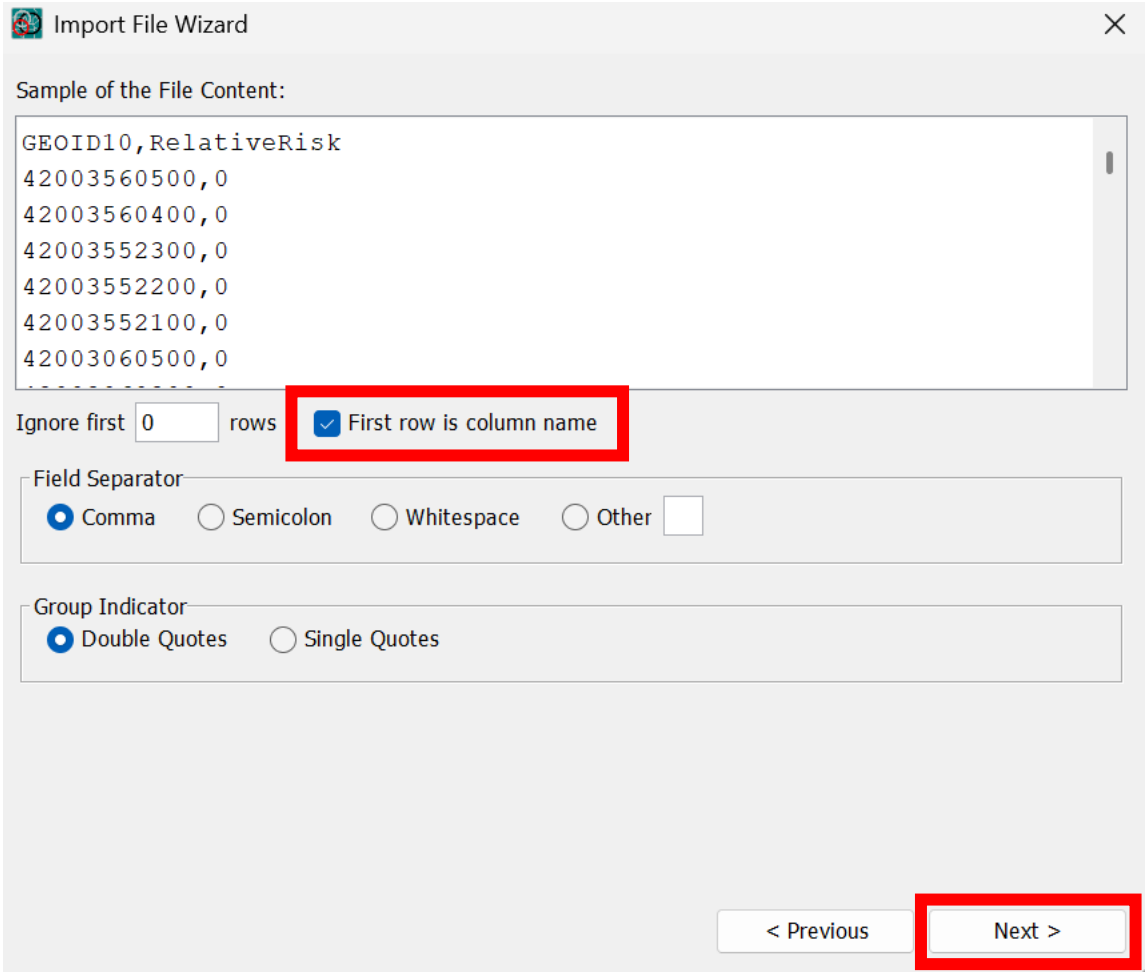
3. Next, navigate to where your data is stored and select the "**PAmiss.csv**" adjustment file that we created for this tutorial. Then choose the "**Open**" button in the bottom right of the window to open it.



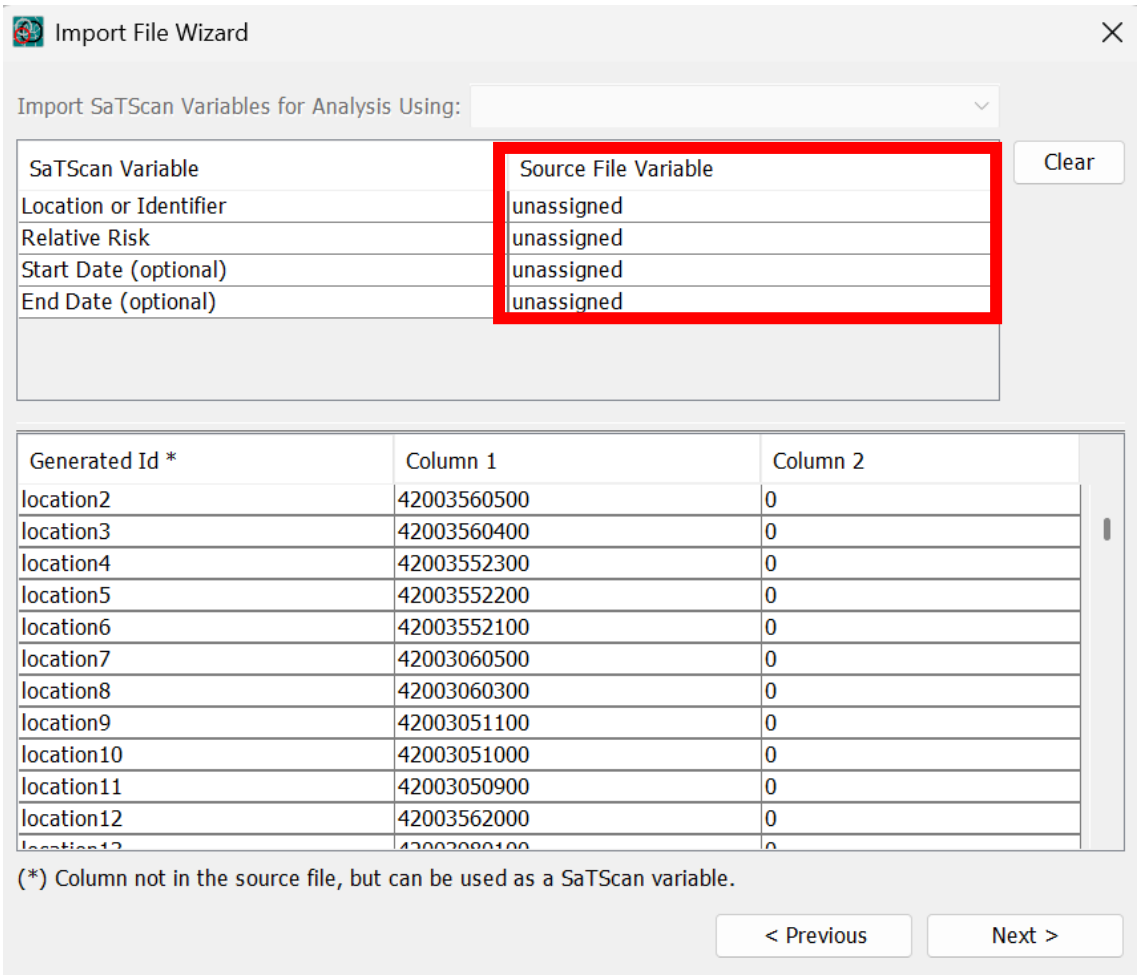
4. The next window below will appear to let you know what the file format should look like. Select the "**Next**" button to continue.



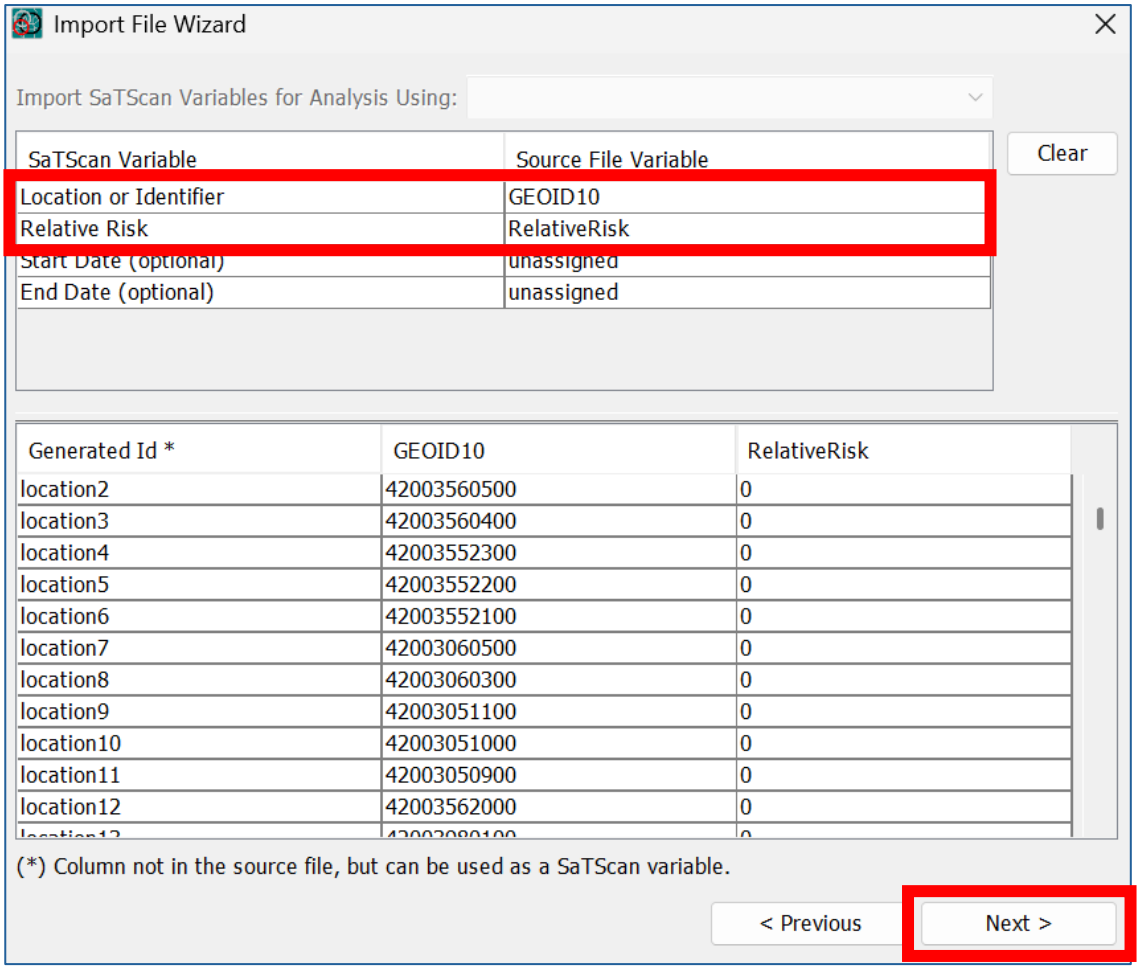
5. You will then get a preview of how SaTScan is reading the data. Select the checkbox next to "**First row is column name**" under the data preview window to let SaTScan know that this file has headers. Then, select "**Next**" to move forward.



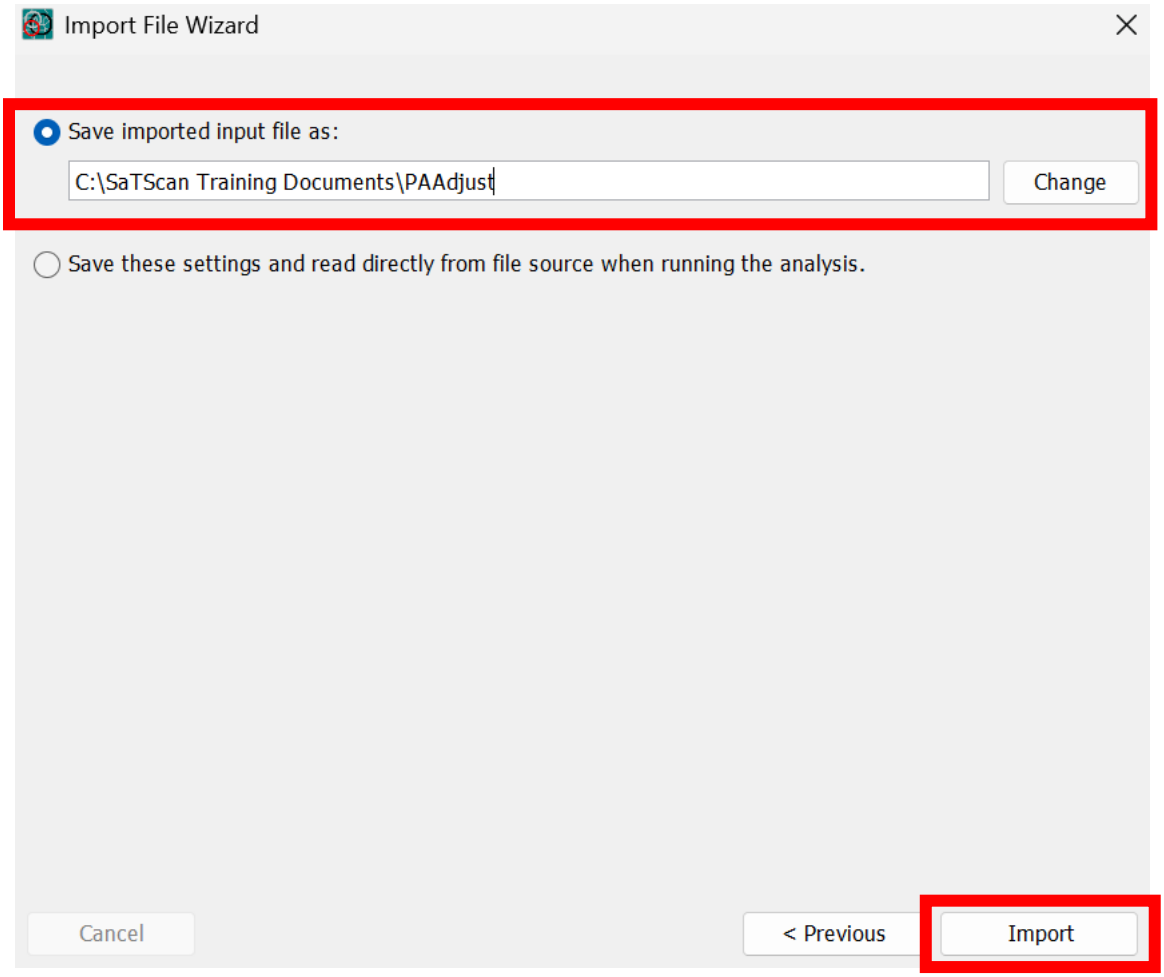
6. Next, assign each required **SaTScan Variable** we will be using for the missing/suppressed data in this analysis to a **Source File Variable**, by clicking on the "**unassigned**" box next to the required SaTScan variables and selecting the appropriate variable from the drop-down menu.



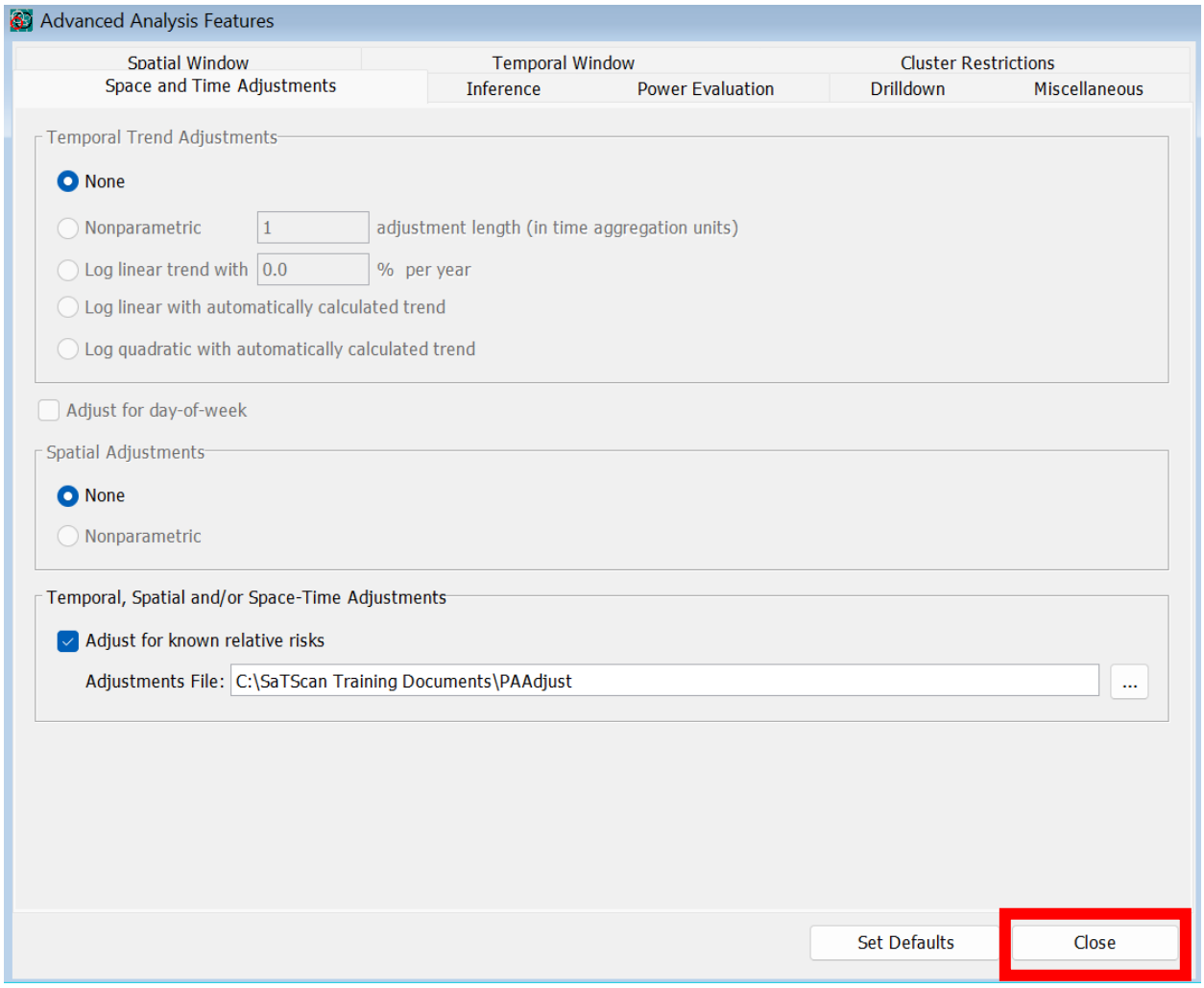
7. For the adjustments file, assign **[Location ID]** to **[Column 1]** and the **[Relative Risk]** to the **[Column 2]** variable in the Source File. Leave the remaining optional SaTScan variables unassigned. Select "**Next**" to move to the next screen.



8. Now, let's turn this adjustments file into a SaTScan file by updating the file location and name in the address bar or by selecting the "**Change**" button to choose a new location. Update the file name in the address bar to "**PAAdjust**" for this tutorial. Select "**Import**" in the bottom right when finished.



9. Lastly, save and close the **Advanced Analysis Options** window, by selecting the "**Close**" button in the bottom right of the screen.

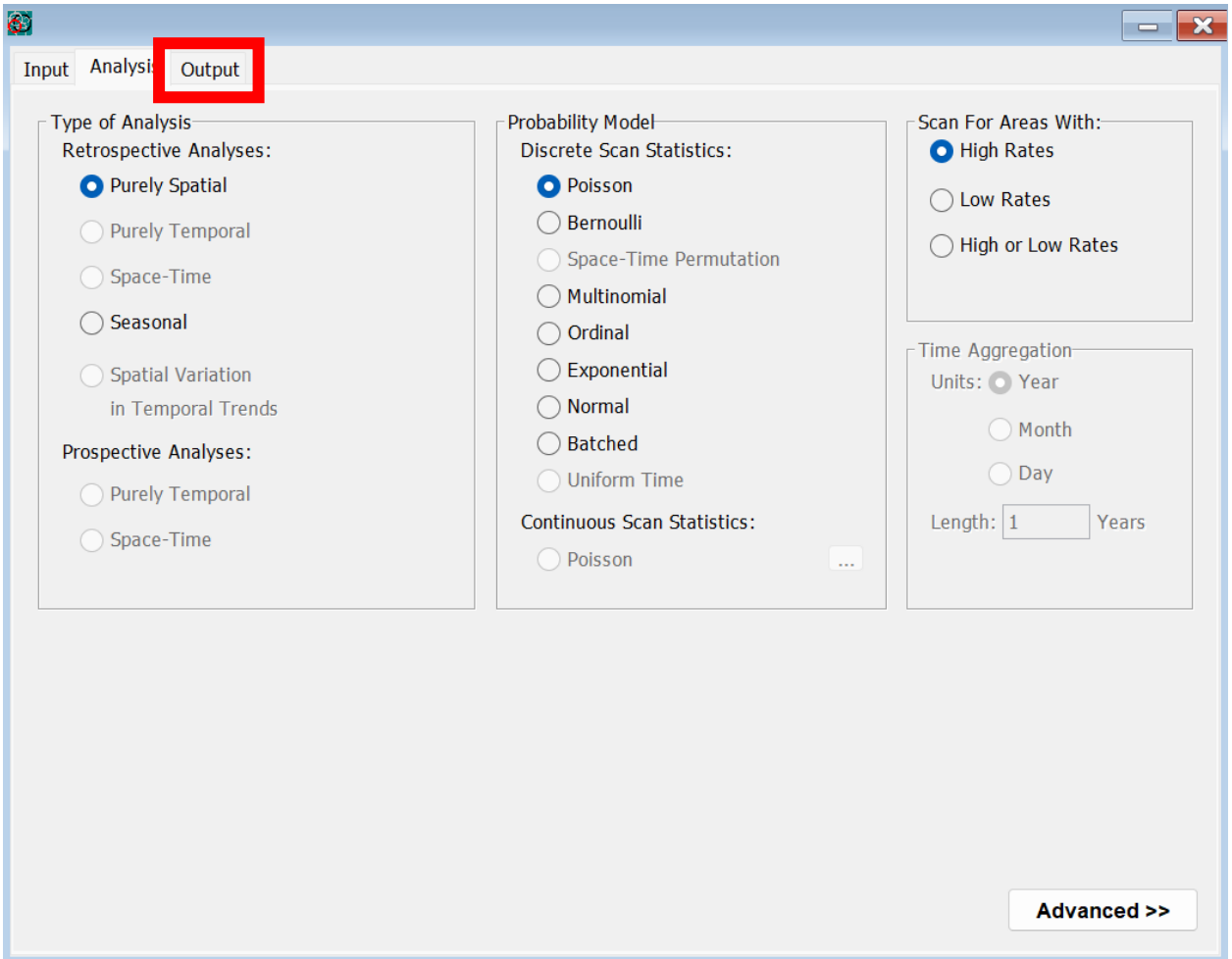


Output

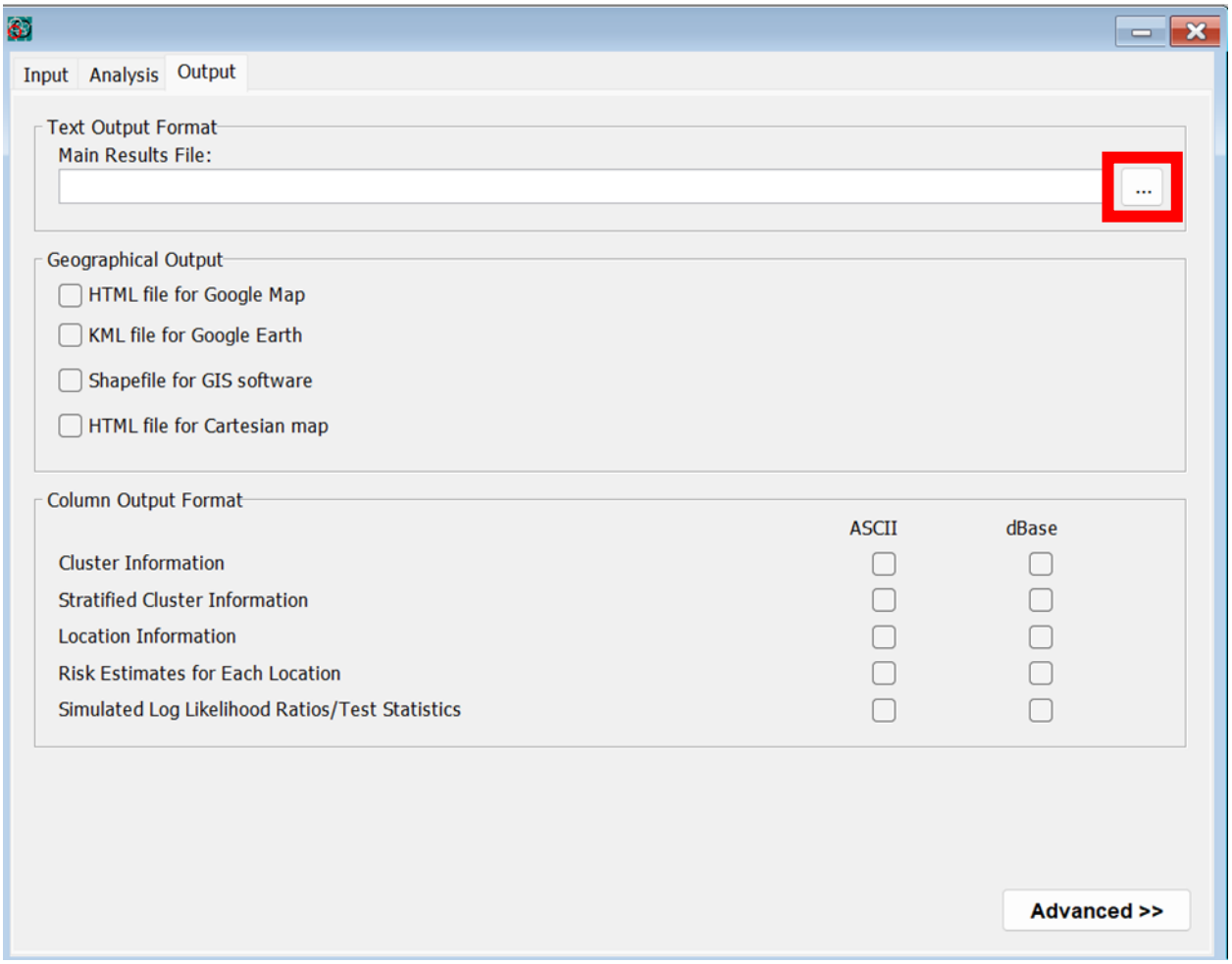
The final step before we run our SaTScan cluster detection analysis is specifying where we want our results to be saved and what format we would like to view them in, such as an HTML file, KML file for Google Earth, or a Shapefile. As HTML files do not require any additional software and/or experience to use, we will focus on viewing the results in an HTML file for Google Map.

***Note**: The SaTScan User Manual in the Help menu contains more information about other ways you can export, review, and summarize the results depending on your needs. You can view the User Manual online (https://www.satscan.org/techdoc.html).*

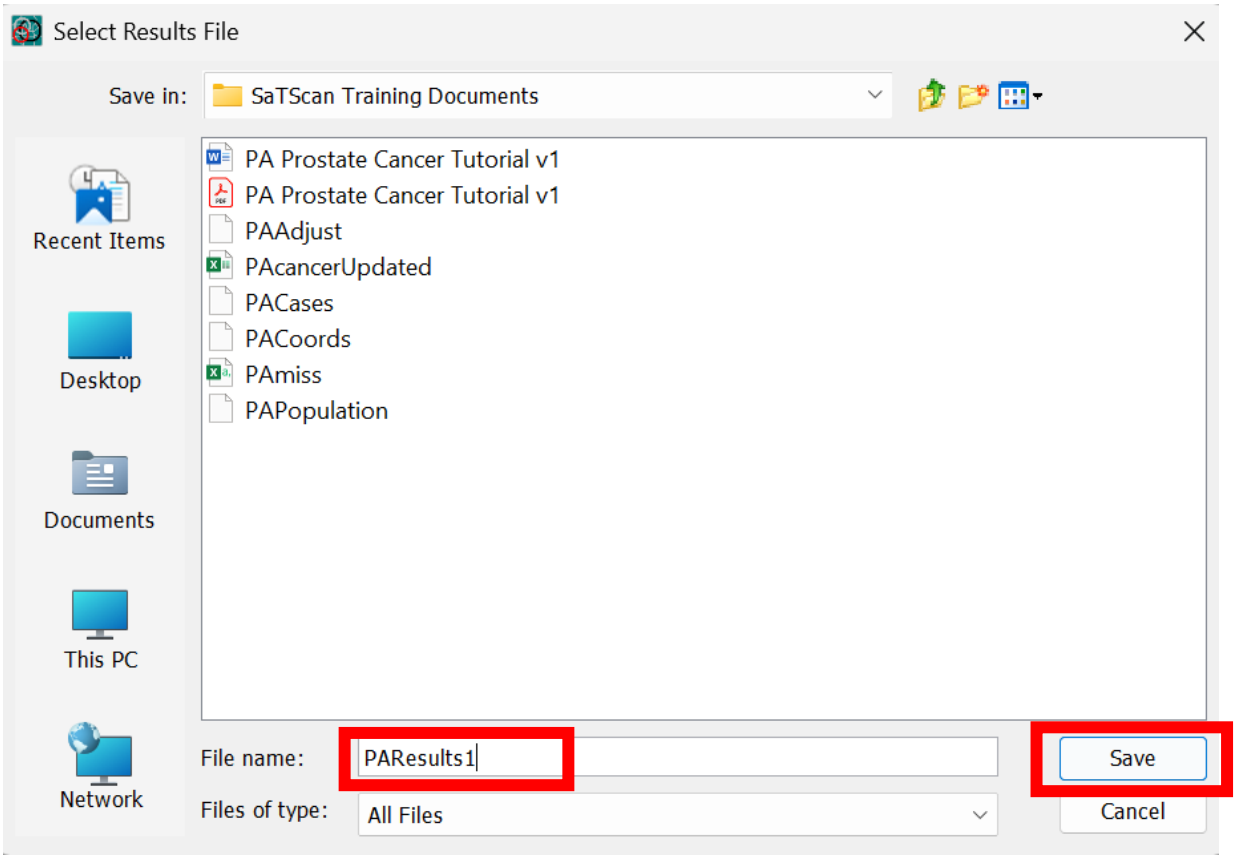
1. First, select the "**Output**" tab at the top of the window.



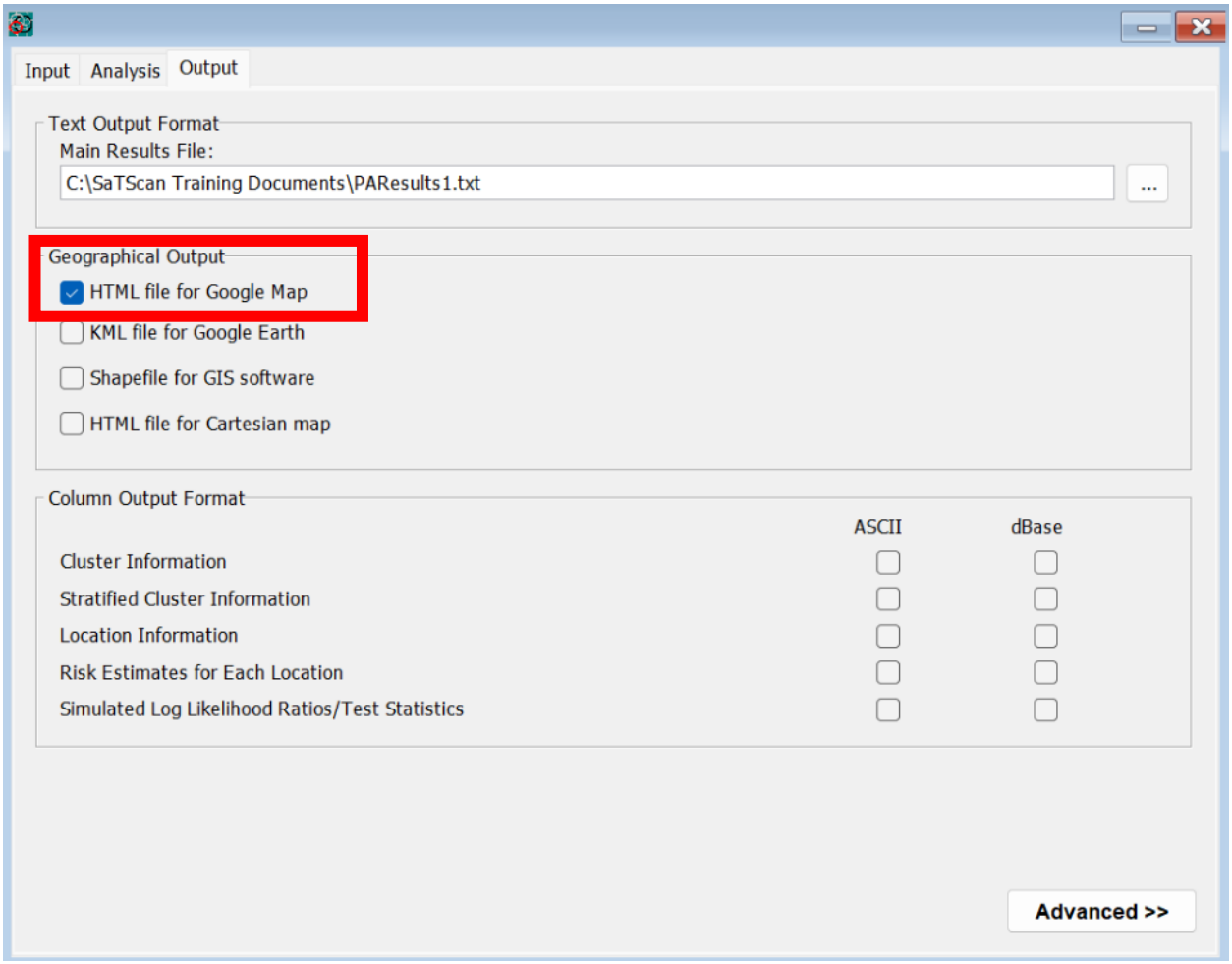
2. Then specify where you would like to save your results by updating the file location and name in the address bar, or by selecting the "**ellipses (…)**" next to the **Main Results File** line.



3. For this tutorial, name the results file "**PAResults1**" and select the "**Save**" button in the bottom right to close the menu.

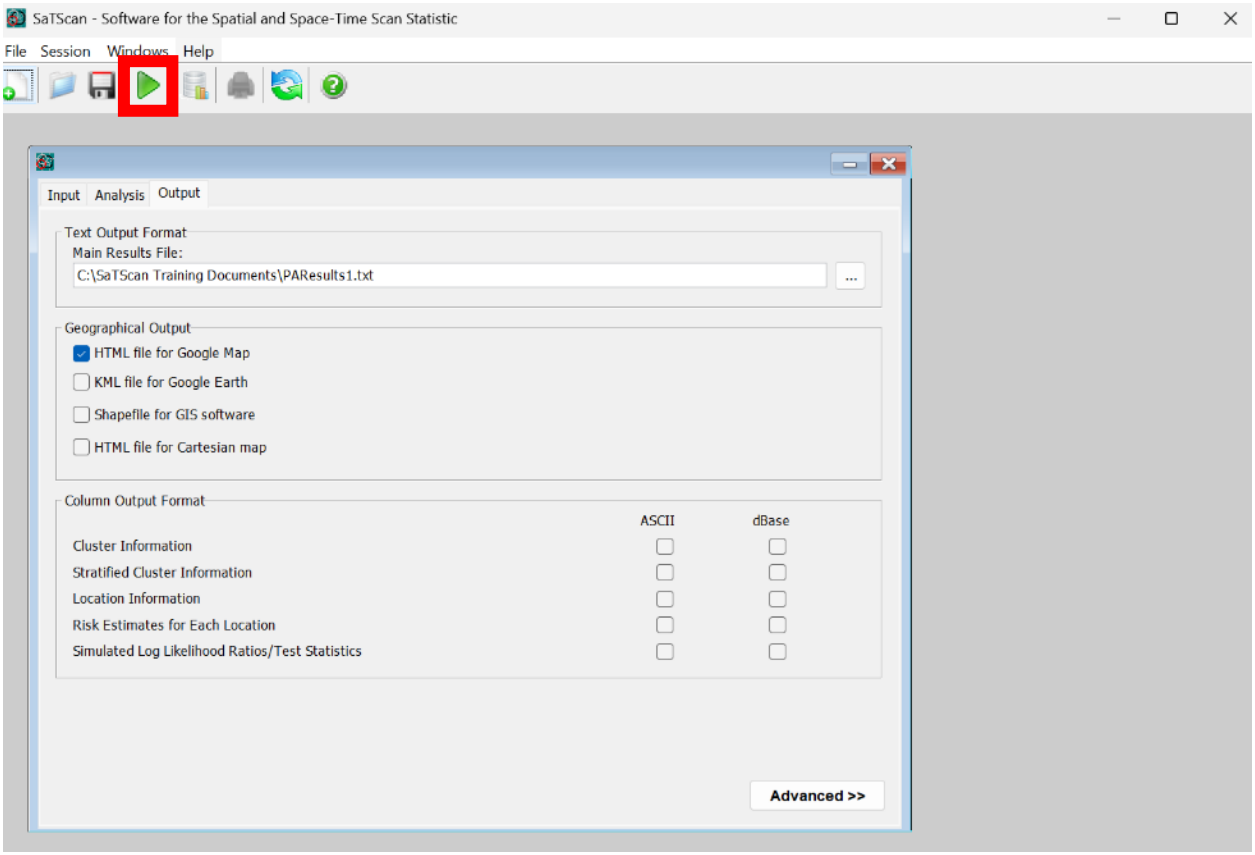


4. We can also have SaTScan automatically open Google Maps so we can quickly visualize the results, by checking the "**HTML file for Google Map**" box in the **Geographical Output** section.



Note: *You can also select additional output options, such as the "Shapefile for GIS Software" option, which will create a shapefile that you can use for more advanced mapping and customization of the results in your favorite GIS software. You can further customize your output, using the "Advanced" button in the bottom right of the "Output" tab.*

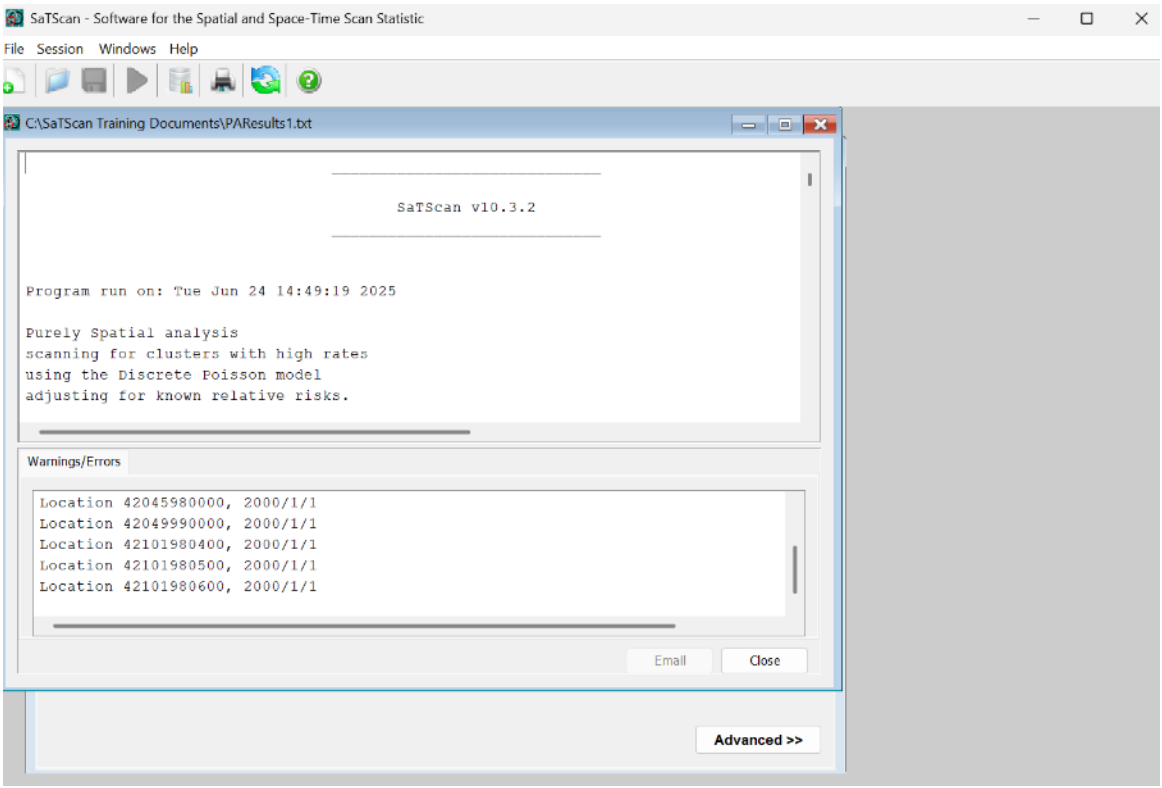
5. Finally, select the "**Run**" button (the green arrow) at the top of the main SaTScan window to run the analysis.



Results

If SaTScan runs without any errors, your screen will match the below image. Otherwise, SaTScan will stop and provide you with specific error messages to try and help you identify what needs to be updated in order to have a successful run.

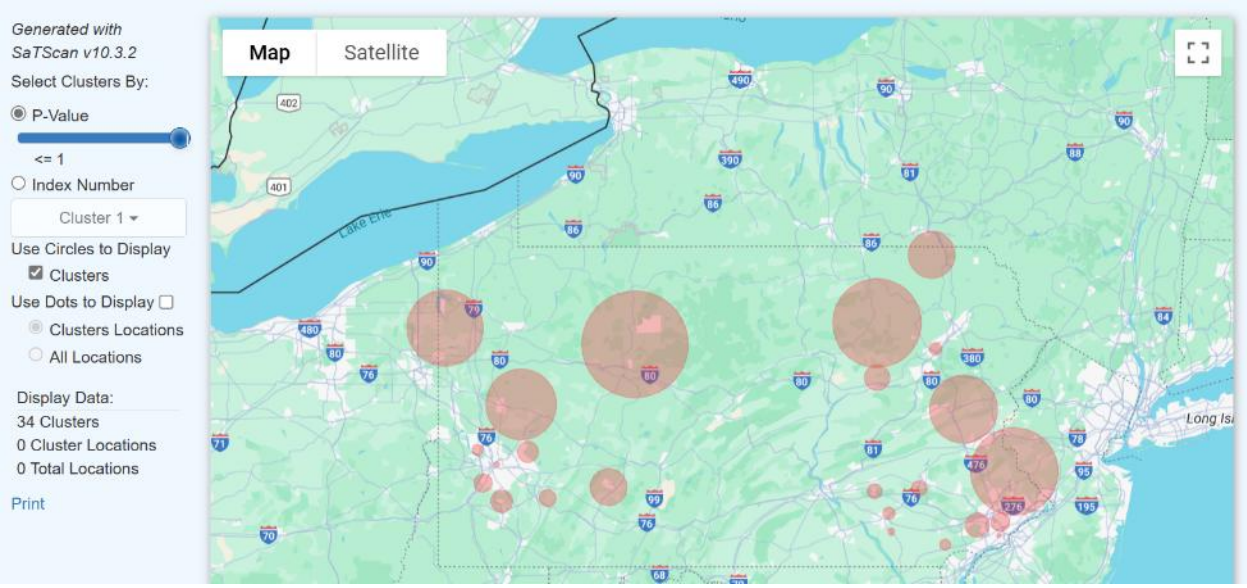
***Note**: Many errors are usually due to data entry errors; double checking that your data and import files are formatted correctly should help solve any issues you run into.*



Mapped Results

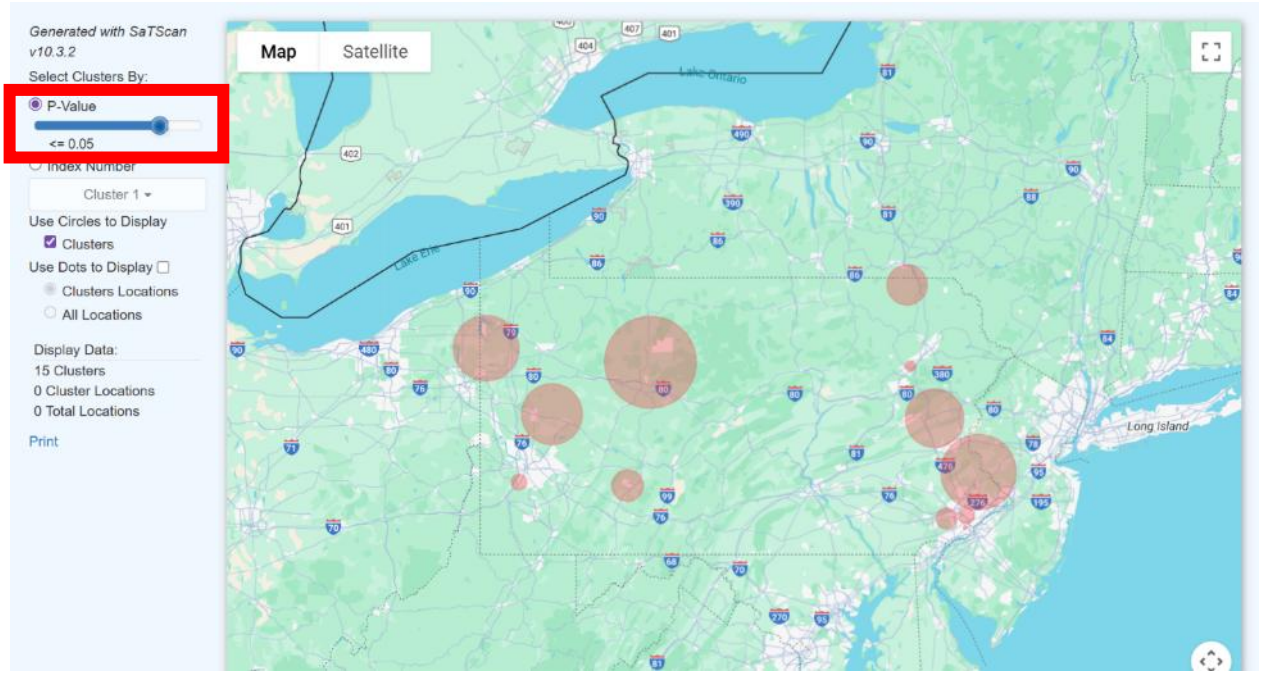
Since we chose to view our results in an "**HTML file for Google Map**", SaTScan will either automatically open the resulting map in Google Earth (if you have it installed) or will let you choose what program on your computer to open it in. If you do not have Google Earth installed, select an internet browser (e.g., Edge, Chrome) to open the results.

Your results will look something like the image below, but there may be some slight differences in functionality depending on your version of SaTScan and what program you are using to view the results. We used SaTScan v10.3.2 and Microsoft Edge to share the results below. Differences in any identified clusters can also result from the Monte Carlo simulation approach, which is used to help determine cluster significance. For more information, see the SaTScan User Guide (https://www.satscan.org/techdoc.html).

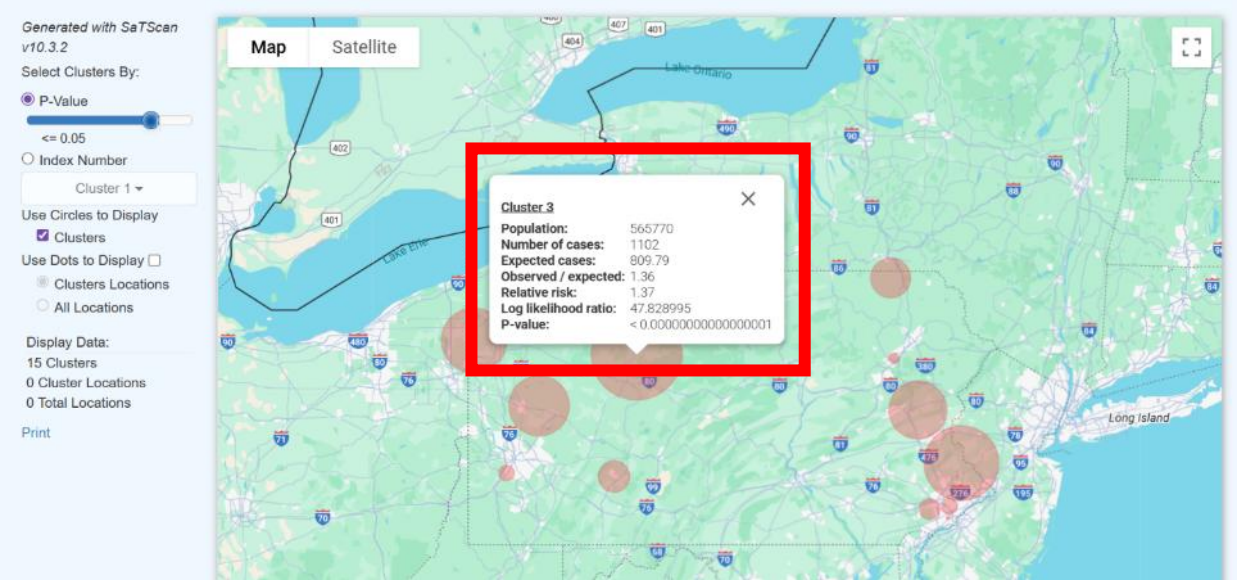


The resulting map is interactive and includes some additional functionality in the left-hand pane, including the ability to filter clusters by significance, learn more about each cluster, and add cluster point locations to the map. Let's test out some of this functionality next.

1. First, let's limit the results in the map, so that we only see **statistically significant clusters (<=0.05**), by dragging the **"P-Value"** bar on the left until you get to "**<=0.05**".

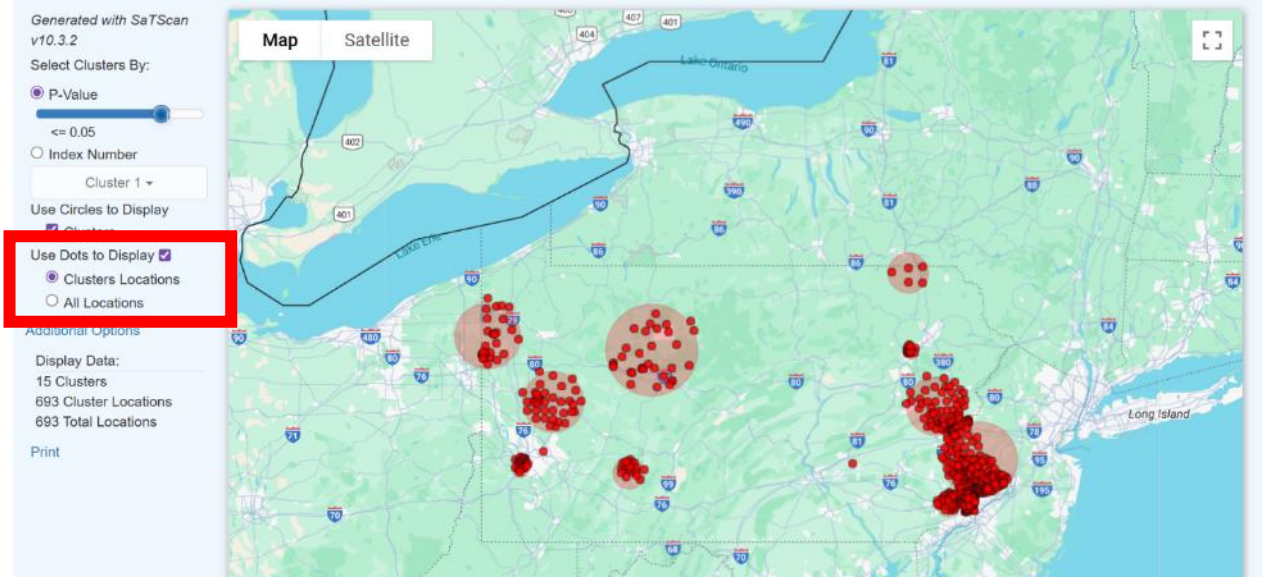


2. Next, select a **cluster** on the map to get more information about it, such as the number of cases and number of expected cases in that cluster. You can also select the "**x**" at the top of the pop-up to close it or select additional clusters to compare values between clusters.



Note: *The SaTScan User's Guide also includes more information on some of the other summary statistics that are shown here for each cluster.*

3. Lastly, let's display the cluster locations as points on the map, by selecting the checkbox after "**Use Dots to Display**" and making sure that **"Cluster Locations"** is selected underneath it. In this case, cluster locations represent the census tract centroids in each cluster.



Results Summary

Before we dive into the results, remember:

This analysis is only intended as a demonstration of how to use SaTScan for cancer spatial cluster detection investigations and the sample results and findings presented as part of this tutorial should not be interpreted as real-world conclusions.

In our example, SaTScan identified **15 significant clusters**, covering **693 census tracts**.

Each cluster represents an area where the number of prostate cancer cases was **significantly higher than expected**, based on the overall PA prostate cancer rate and population. Some clusters were just a single tract, while others included several tracts. Remember, we set the maximum cluster size, to make sure that no cluster would include more than **10% of the total male population (18+ years) in PA**, which helped keep any one cluster from getting too large.

Note: If your analysis did find clustering present, it does not necessarily mean there is a single cause or environmental cause for the pattern; it is information to consider for further investigation. Similarly, if your analysis did not find clustering, it does not necessarily mean that analyses are complete. Investigators can use findings from these analyses to inform future work and investigations.

BONUS: Age-Adjusted Analysis Tutorial

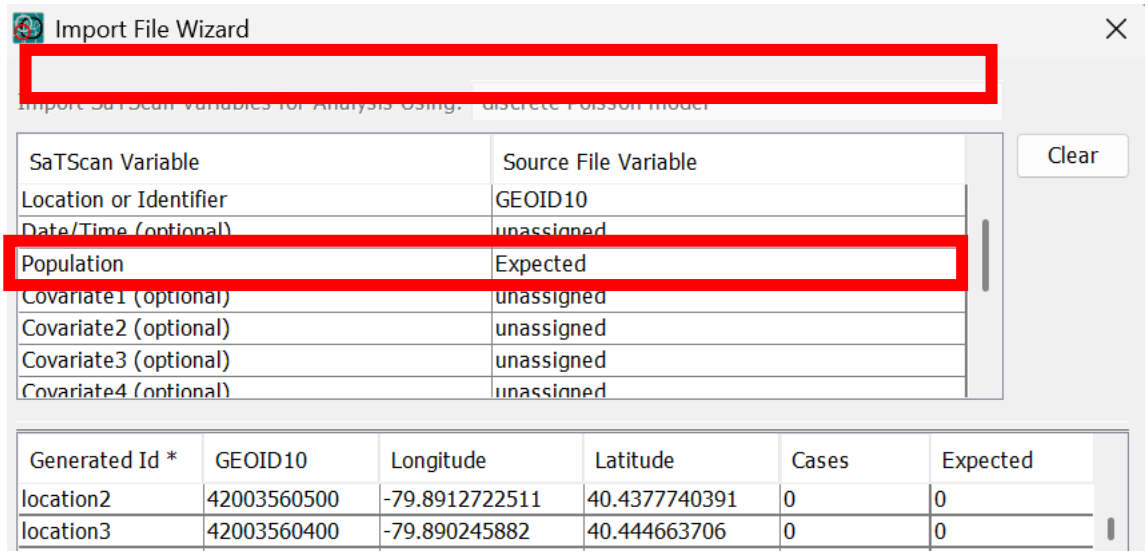
The SaTScan cluster detection analysis we just ran is often referred to as an **unadjusted analysis**, as we did not use or account for characteristics of the underlying population (e.g., age) that may affect the number of cancer cases we would expect in each tract. However, most cancers have a known **age effect** (i.e., the risk of that cancer increasing with age), including prostate cancer. It can be useful to run an **age-adjusted analysis** for these cancers, to help account for areas where we would naturally expect more cancer cases based on the age of the population living there.

For example, imagine that one of the clusters we identified in the unadjusted analysis is located in an area with more older men than anywhere else in the state. While our original analysis may have identified a statistically significant cancer cluster in that area, this cluster might disappear if we ran an age-adjusted analysis, as we would naturally expect to see more prostate cancer cases in parts of the state where more older men live.

Age-Adjusted Analysis Tutorial

To perform an age-adjusted analysis in SaTScan, we will need: (1) age information for all cancer cases, and (2) age information for the entire population in the study area. We have included this information in our **PAcancerUpdated.xlsx** file for today's tutorial. Run the age-adjusted analysis by repeating the unadjusted analysis with the following changes:

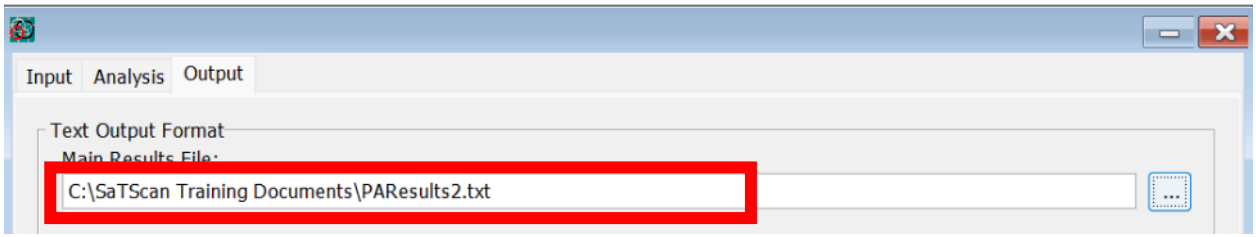
1. First, go back to the **Import** tab. Leave everything in the **Case** and **Coordinates** files the same. For the **Population Data** file, leave everything the same as before, except for the [**Population**] variable. Instead, select the [**Expected**] variable from our source file from the drop-down menu, which represents the age-adjusted expected counts of prostate cancer per census tract.



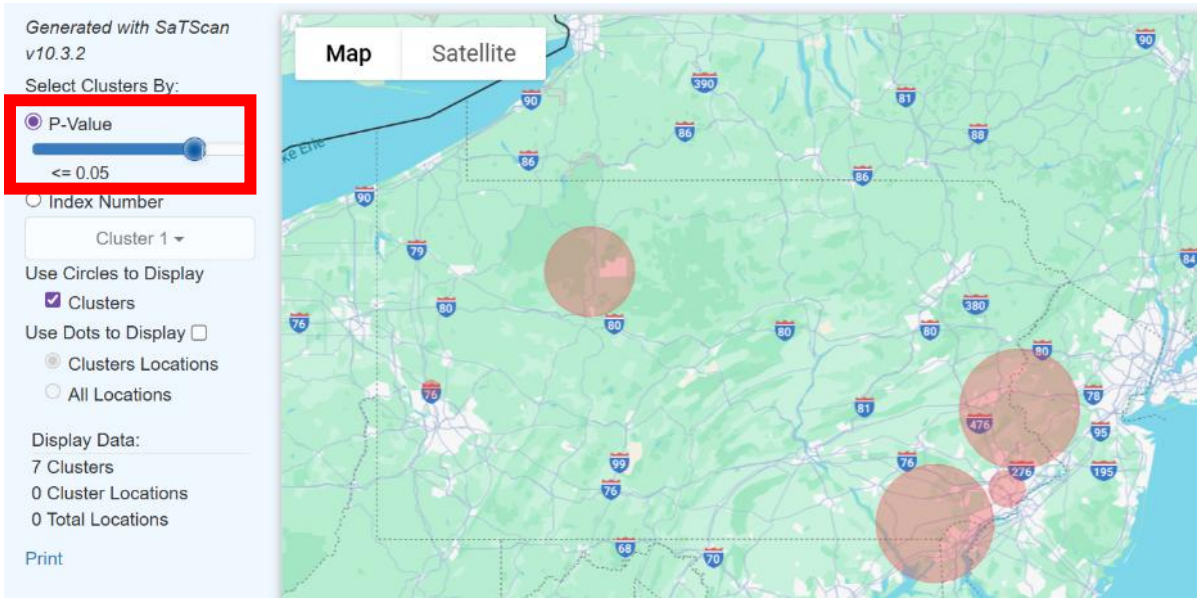
Save the updated **Population** file as **PAPopulation2**, to make sure we can tell the difference between the unadjusted and age-adjusted population files.

2. Next, make sure that we use the same settings for the **Analysis** tab that we used for our unadjusted analysis.

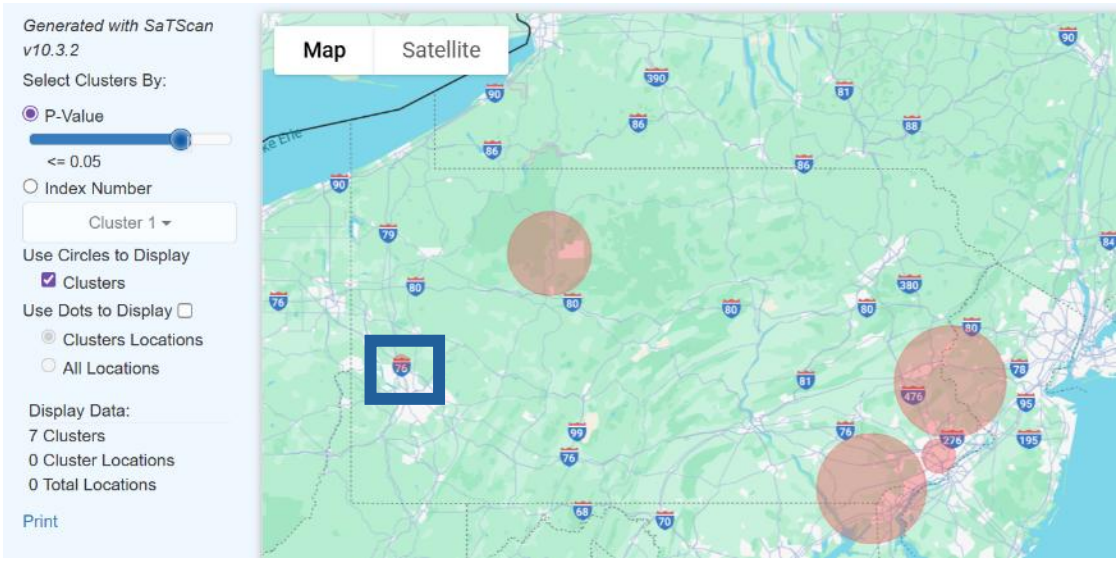
3. For the **Output** tab, we will want to rename our new age-adjusted results to **PAResults2**, to help differentiate between our unadjusted and age-adjusted results.



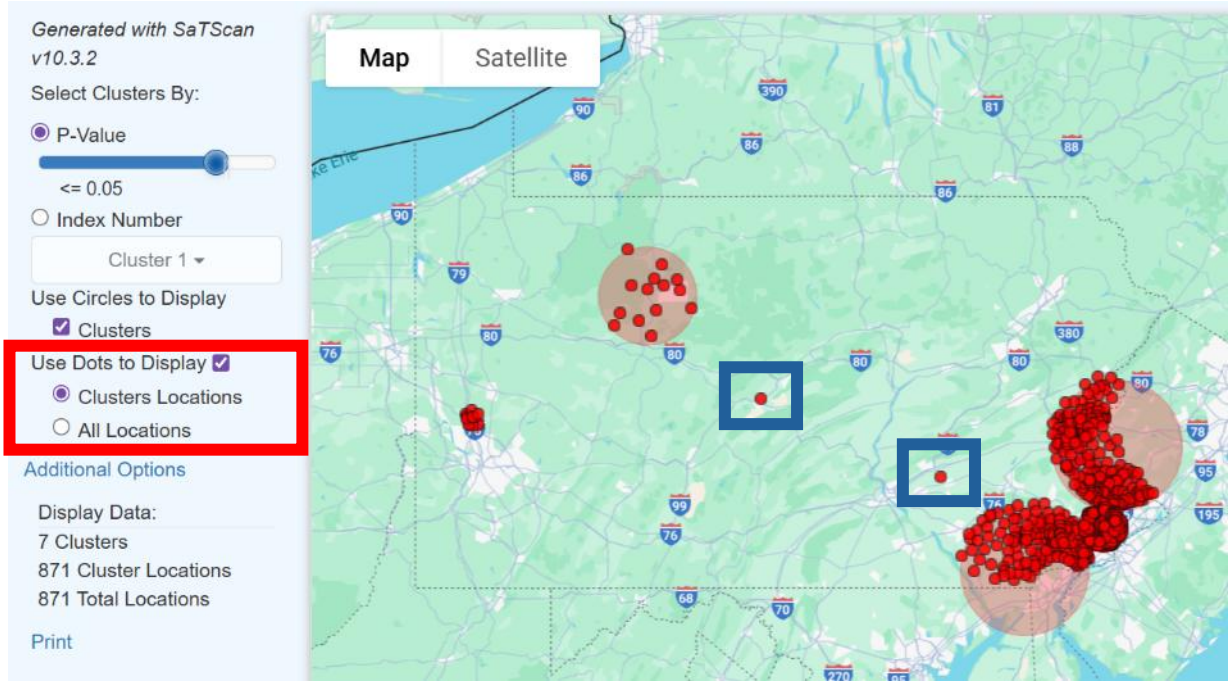
4. **Run** the analysis using the green arrow in the main menu (▶) and adjust the **P-Value scale bar** to show only statistically significant clusters on the map.



5. There are 7 statistically significant clusters in the age-adjusted results, including one smaller cluster that overlaps the I-76 sign on the left side of the map.



There are also two additional statistically significant clusters in central PA that are too small to see at this scale. You can use the **"Use Dots to Display"** feature to see these smaller clusters on the map or zoom in until the cluster circles appear.



Let's compare the results of our age-adjusted analysis to the results from our unadjusted analysis next.

Comparison Summary

Comparing the results of our adjusted and unadjusted analyses can be a helpful way to better understand and explain patterns in cancer data. Let's compare the results of our **unadjusted** (Figure 1) and **age-adjusted** (Figure 2) analyses. When we adjusted the analysis for age, the number of statistically **significant clusters** dropped from **15** (in our unadjusted analysis) to **7** (in our age-adjusted analysis). This means that **6** of the **statistically significant** clusters that we saw earlier, can be explained by variations in age across PA census tracts. The remaining **7** clusters we see in our age-adjusted analysis results are **clusters that cannot be explained by differences in age alone**.

When you adjust for factors (like age), the shape and number of clusters can also shift. We see this in our results as well, with how many of the clusters in Western and Northern PA disappear, while the clusters in the Southeastern part of the state grow much larger, when we adjust for age.

Figure 1 – Unadjusted Results

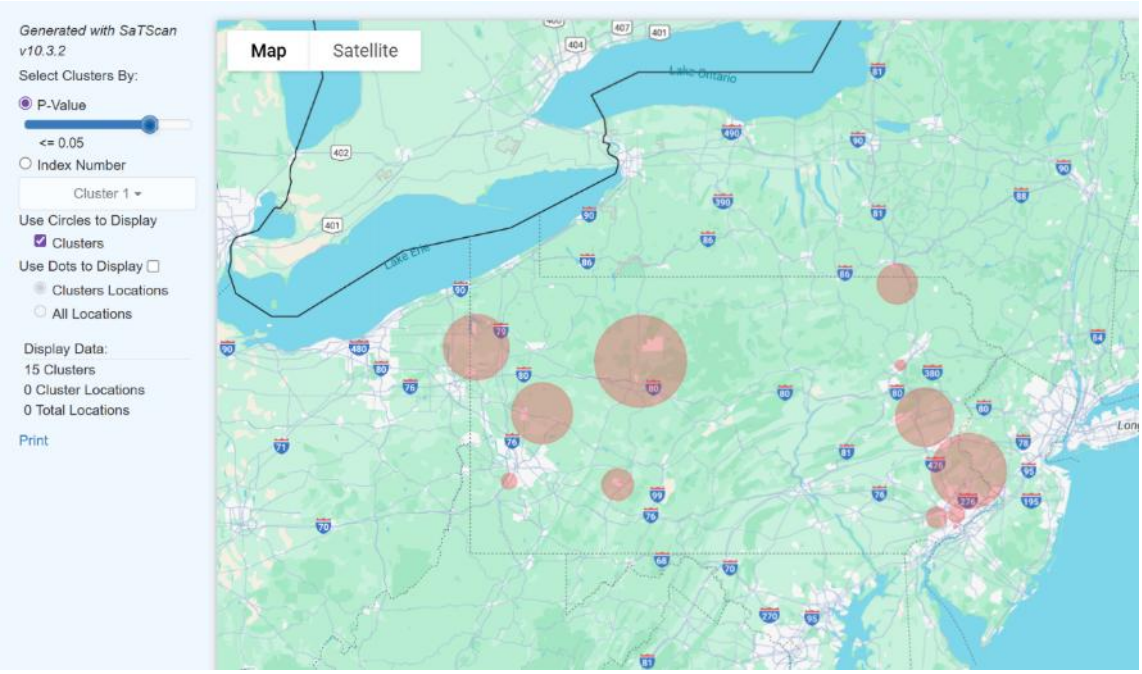
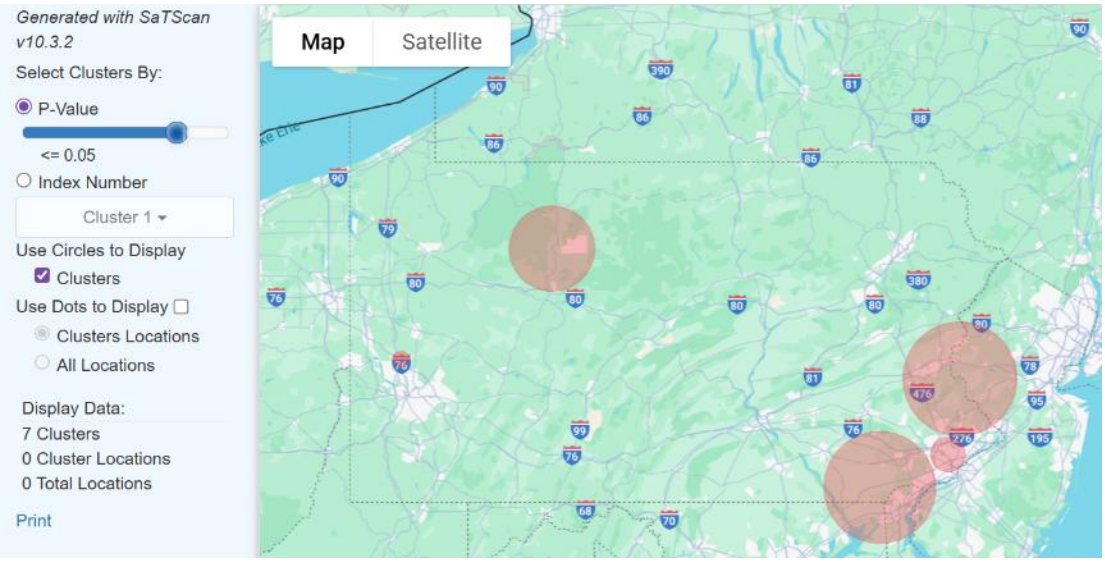


Figure 2 – Age-adjusted Results



To learn more about how SaTScan handles adjustments like those seen in this tutorial demonstration, see the SaTScan User's Guide.

References

SatScan. Accessed: https://www.satscan.org/.

SatScan Technical Documentation. Accessed: https://www.satscan.org/techdoc.html.

SatScan Tutorials: Accessed: https://www.satscan.org/tutorials.html.

Pennsylvania, Age-Adjusted Rate of Prostate Cancer (Males Only) per 100,000 Population, Census Tract, 2010-2019 - https://ephtracking.cdc.gov/DataExplorer/?query=ab3f12d9-49b6-4e74-b6fe-e2d80ff0cb5d.

CDC, Centers for Disease Control and Prevention. Guidelines for Examining Unusual Patterns of Cancer and Environmental Concerns. December 2022. Accessed: https://www.cdc.gov/cancer-environment/php/guidelines/index.html.