Introduction to Outlier Detection

A SpaceStat Software Tutorial

Principal Investigators: Pierre Goovaerts and Geoffrey Jacquez . SpaceStat Team: Eve Do, Pierre Goovaerts, Sue Hinton, Geoff Jacquez, Andy Kaufmann, Sharon Matthews, Susan Maxwell, Kristin Michael, Yanna Pallicaris, Jawaid Rasul, and Robert Rommel.

SpaceStat was supported by grant CA92669 from the **National Cancer Institute** (NCI) and grant ES10220 from the **National Institute for Environmental Health Sciences** (NIEHS) to BioMedware, Inc. The software and help contents are solely the responsibility of the authors and do not necessarily represent the official views of the NCI or NIEHS.

BioMedware
Geospatial Research and Software

MATERIALS

Outlier Detection_Tutorial.SPT SpaceStat project

ESTIMATED TIME

30 minutes

OBJECTIVE

This tutorial will conduct an exploratory data analysis to identify statistical and spatial outliers in SpaceStat. The data are age-standardized lung cancer mortality rates from 1950 to 1995. In this exploratory analysis you will use boxplots, variogram clouds, and the local Moran to detect outliers.

WHAT ARE OUTLIERS?

Outliers refer to data points that are different from the other data. There are two types of outliers that we will look at in this tutorial. There are statistical outliers where the value is unlike the values in the rest of the distribution. These data points are either higher or lower than the other data points. There is another type of outlier, spatial outliers, where data points are different from the data points around them. These points may not be extremes in the entire distribution but are so when you consider their local geographic neighborhood.

ABOUT THE DATA

The data used for this tutorial are lung cancer mortality rates across the United States at the county level. The data come from the National Cancer Institute's National Atlas of Cancer Mortality and are age-standardized mortality rates per 100,000. The population we are looking at in this exercise is white males and the data spans from 1950 to 1995. Further details on the data can be found at http://www3.cancer.gov/atlasplus.

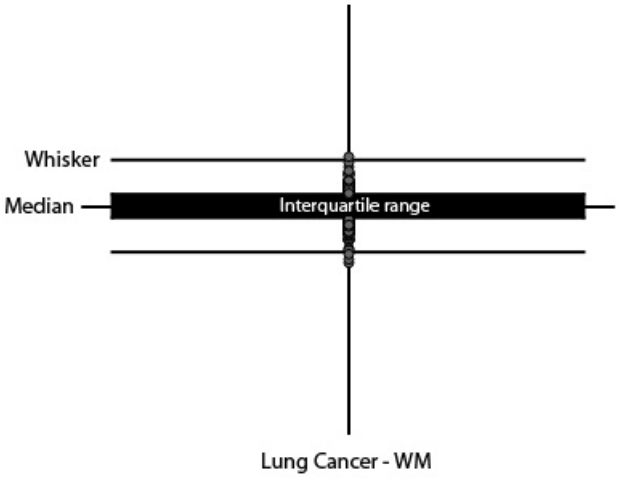
STEP 1: LOAD THE PROJECT

To open this project, **go to File -> Open**, and then **browse to "Outlier_Detection_Tutorial"**. **Click the "OK" button** and the project will load. Look at the Datasets listed in the Data view. You will notice that the datasets are listed under the 'Counties' geography. The datasets have been renamed from the original files from the National Cancer Institute. There are datasets for both the rate and counts of white male lung cancer mortality. The rates represent the age-adjusted mortality rates per 100,000. The 'Lower Bounds' and 'Upper Bounds' datasets contain lower bounds of the 95% confidence interval of the rate.

STEP 2: FINDING STATISTICAL OUTLIERS WITH BOXPLOTS

Click the Boxplot button ( **)** on the main SpaceStat toolbar. **Select 'Counties'** in the Geography dropdown control. **Select 'Lung Cancer - WM'** in the dataset dropdown control. **Click the 'OK' button** to create the boxplot.

The boxplot is a unique visualization. Data observations are shown as points along a vertical line. There is horizontal black-filled box that is usually towards the center of the plot. This box is the 'interquartile range'; this range contains 50% of the values, specifically the 25% above the median and 25% below the median. A black line extends through the center of this box, representing the median value. There are also two additional lines on the boxplot above and below the interquartile range called 'whiskers'. These show values +/- 2 times the interquartile range from the median (remember that the interquartile range is bounded by values +/- 0.5 times the interquartile range).



A boxplot with labeled parts

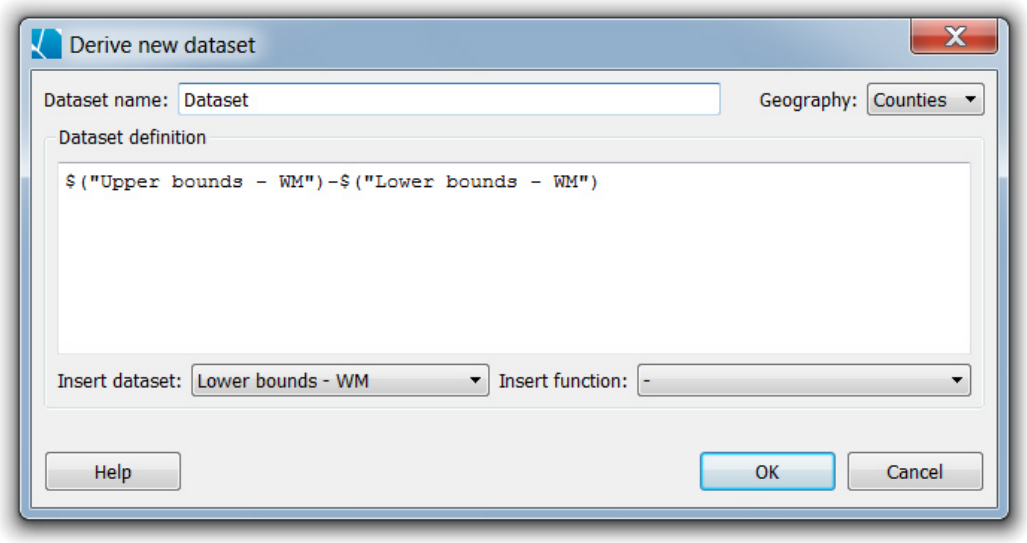
Any values that fall outside the interquartile range are candidates to be considered outliers. For instance, there are a few values that fall above the top whisker in 1950.

Create a map of the Lung Cancer – WM dataset by clicking the 'Map' button from the main toolbar. Select the 'Lung Cancer – WM' as the Fill Color Dataset. With both the map and the boxplot open, drag a selection rectangle around points on the boxplot. This 'brush' selection can be used to show where possible outliers are on the map. Advance the time slider to 1970 and brush select the possible outliers at this time. At this stage in an exploratory analysis you might try to think of possible explanations for why these values are higher or lower than the others.

STEP 3: CREATING AN UNCERTAINTY DATASET

The data contains two variables for each population that can be used to create a measurement of uncertainty. Two of the datasets represent the upper and lower bounds of the 95% confidence interval around the white male lung cancer mortality observations. The difference between these two values is a measure of the uncertainty of the lung cancer mortality rate.

SpaceStat can create new datasets from existing data using the Derive New Dataset operation. Start the procedure by **selecting 'Tools->Derive new dataset'** from the SpaceStat main menu. **Name the new dataset 'uncertWhiteMale'**. **Select 'Counties'** from the geography dropdown. Next **choose 'Upper Bounds - WM'** from the Insert Dataset dropdown. Then **choose the '-'** from the Insert Function dropdown. **Choose 'Lower Bounds - WM'** from the Insert Dataset dropdown. The dialog should look like the following screenshot (pay special attention to the dataset definition):



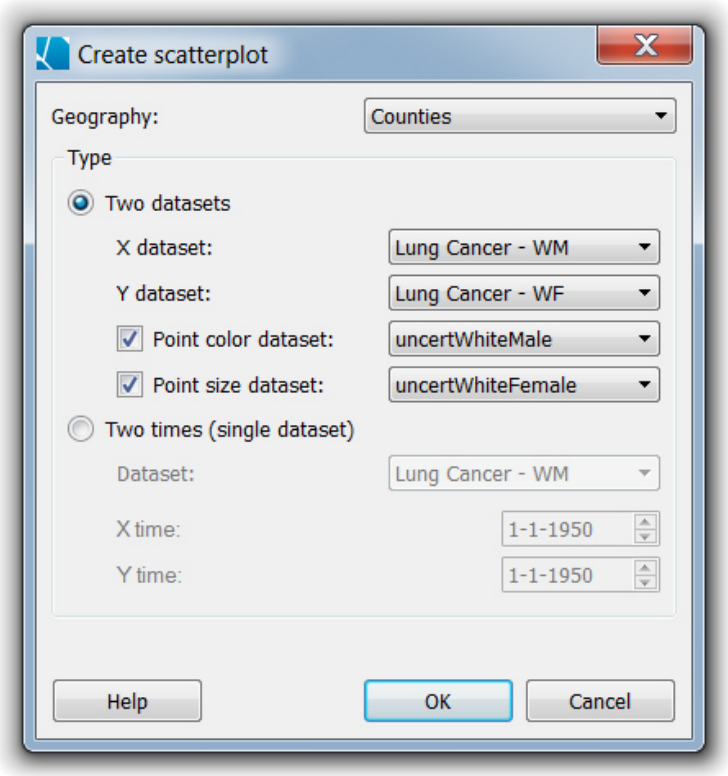
Creating an uncertainty dataset

If your dialog matches this screenshot, **click the 'OK' button** to create the dataset. If you have made an error, you can either edit the text manually or delete the definition and start over. **Repeat this process to create an uncertWhiteFemale dataset from the 'Upper Bounds – WF' and 'Lower Bounds – WF' datasets**.

STEP 4: MULTIVARIATE RELATIONSHIPS

The scatterplot tool in SpaceStat can create a graph of one variable against the other and will fit a line through it showing the correlation coefficient. It is a good starting point to explore a bivariate relationship.

Click the Scatterplot button (**)** on the SpaceStat main toolbar. **Select 'Counties'** from the Geography dropdown control. Make sure that the type of scatterplot is set to 'Two datasets'. **Select 'Lung Cancer - WM'** from the X dataset dropdown control. **Select 'Lung Cancer - WF'** from the Y dataset dropdown control. Check the 'Point color dataset' option and choose 'uncertWhiteMale' from the dropdown. Check the 'Point size dataset' option and choose 'uncertWhieFemale' from the dropdown. Check to make sure your dialog matches the screenshot below and **click the 'OK' button** to create the scatterplot.



Create scatterplot dialog settings

The county data was collected at two points in time. If you advance the slider forward to 1970, you should see the data change. Some things you should consider at this point are whether there is a relationship between the two variables, how strong is the relationship, does the relationship change over time, and does the linear best fit line match the shape of the relationship. 'Bivariate outliers' is a term that is used to describe locations that don't seem to fit the overall relationship between the two variables. These outliers would appear far away from the best fit line in the scatterplot. Note that in this dataset there are some locations for which the female or male lung cancer rate (the female is more common) is recorded as 0 due to a small population. How do you think these observations impact the bivariate relationship?

STEP 5: IDENTIFYING SPATIAL OUTLIERS WITH THE VARIOGRAM CLOUD

The variogram cloud is a visualization that displays the dissimilarity between any two locations as a function of their geographical distance. To be able to use a variogram cloud, you must have a measure of distance. The counties are polygons and as such there is no clearly defined distance measure. **Click on the 'Counties' geography** in the Data View. **Right click this heading and choose 'Create centroid geography'**. This will create a point geography from the polygons with a single point for each county at its centroid. **Make sure the 'Include parent geography's data sets' option is checked.**

Now you can create a variogram cloud. **Click the 'Variogram cloud' button ( ) on the main SpaceStat toolbar.** 'Counties centroids' should be selected as the geography. **Select 'Lung Cancer – WM' as the dataset. Click the 'OK' button** to create the variogram cloud. It may take a few minutes to create the variogram cloud as SpaceStat is calculating values for each possible pair of locations. Your variogram cloud should look like the screenshot below:

Dd

Variogram cloud

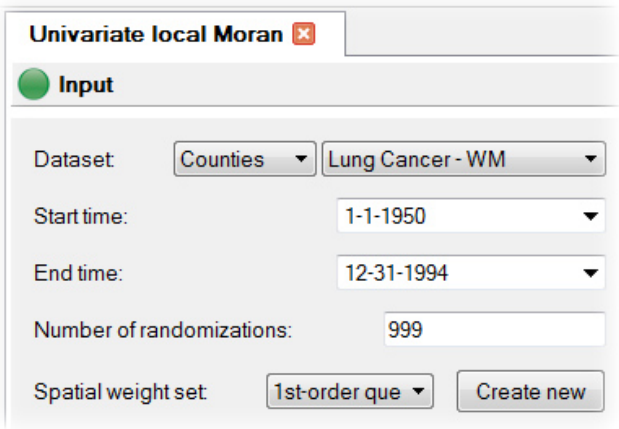
Note that with this many points, SpaceStat displays the results in bins (2D-histogram) rather than a separate point for each data pair. **Create a map of the 'County centroids' geography.** With the map visible, you can select on the variogram cloud and see the centroid pairs that get selected. Remember that the variogram cloud shows pairs of points. Spatial outliers will be those pairs of

points that are dissimilar but relatively close together, i.e. points near the top left of the variogram cloud. If you select these values in 1950, you may notice that many of the point pairs are in the Appalachian region. You can advance the time to the end of the dataset and do the same exercise to see if the spatial outliers change with time. Also you can compare these spatial outliers with the statistical outliers that you saw from the boxplot.

STEP 6: PATTERN AND OUTLIER DETECTION WITH THE LOCAL MORAN

The Moran scatterplot plots the relationship of each observation with the average value of its neighbors. This is useful to examine spatial autocorrelation. High spatial autocorrelation means that locations are highly correlated with their neighbors while no or low spatial autocorrelation means than the value at a location is unrelated to the value of its neighbors.

Run the Moran method in SpaceStat by **selecting Methods->Clustering->Local Moran->Univariate local Moran** from the SpaceStat main menu. This will open the Task Manager pane for you to choose the settings for the method. For the dataset, **select the 'Counties' geography** for the first dropdown. **Select 'Lung Cancer - WM'** for the second dataset dropdown. The default times spanning the whole dataset are fine as is the default spatial weight set. At this point your settings should look like the following screenshot.



Settings for the Univariate local Moran method

Click the Advanced tab and **select 'No'** for the Use Simes correction dropdown control. **Click the Output tab** to review where the results will be stored. Now **click the 'Run' button** to start the method.

When the method finishes, results are placed in several locations. First, the global measure, Moran's I, is reported in the Log View. The Log View is normally at the bottom of the SpaceStat interface, but you may have to enable it through 'Window->Log View' in the SpaceStat main menu if

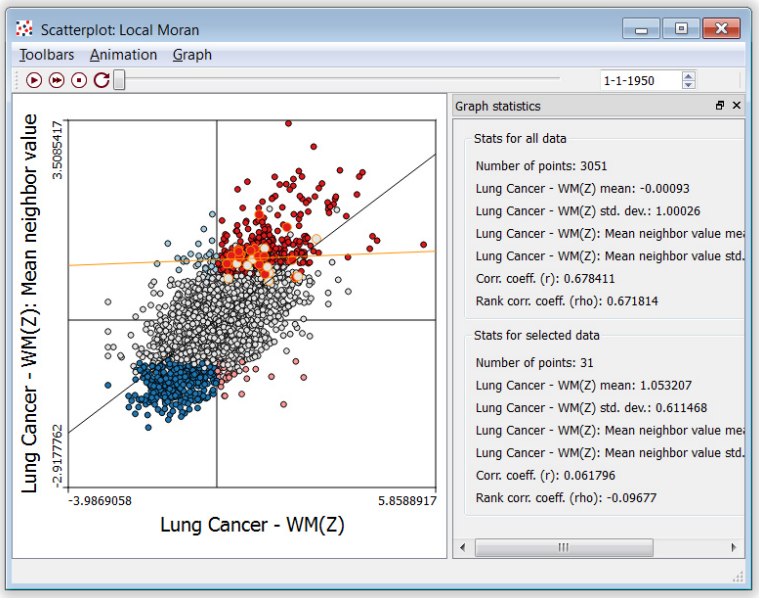
you have closed it. Moran's I ranges from about -1 to 1 (it can be outside this range sometimes due to spatial weights) where 1 is perfect positive spatial autocorrelation, 0 is no spatial autocorrelation and -1 is perfect negative spatial autocorrelation. SpaceStat reports Moran's I for each time period and a p value indicating the probability of obtaining this result under randomization if there was no spatial autocorrelation.

Look at the values reported in the Log View and determine if there is strong or weak autocorrelation in the lung cancer values. Is the spatial autocorrelation positive or negative? There are only two time periods for this data, but does the spatial autocorrelation increase or decrease through time?

Time	Moran's I	P value
[1 Jan 1950, 1 Jan 1970)	0.494698	0.001000
[1 Jan 1970, 31 Dec 1994]	0.678154	0.001000

Table of Moran's I results from the Log View

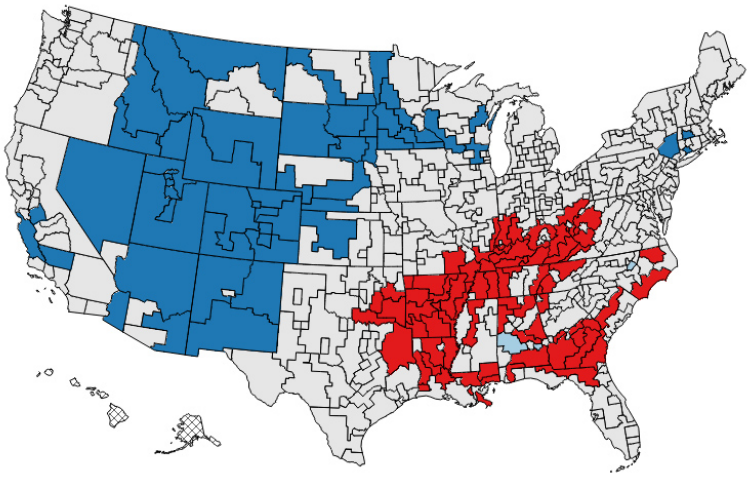
The Moran method also creates a Moran scatterplot. The Moran scatterplot can be used to find spatial outliers. Spatial outliers are those locations that have values very different from their neighbors. The x-axis on the Moran scatterplot shows the rate of the subject SEA while the y-axis shows the average rate of the neighbors. Spatial outliers will be those values that tend to be located in the top left or the bottom right of the scatterplot. Remember that you can brush select these values to see where locations are on the map.



Moran scatterplot

Lastly the Moran method creates a map of the locations with locations classified by their value and the value of their neighbors. There are 5 possible classes. The first class is locations that are not

significant clusters or outliers. There are two types of significant clusters: High-High and Low-Low. High-High (shown in SpaceStat in red) corresponds to locations that have a high value and are surrounded by other locations with high values. Low-Low (shown in SpaceStat in dark blue) corresponds to locations that have a low value and are surrounded by other locations with low values. There are also two types of significant outliers: High-Low and Low-High. High-Low (shown in SpaceStat in pink) corresponds to locations with values that are significantly higher than the values of their neighbors. Low-High (shown in SpaceStat in light blue) corresponds to locations with values that are significantly lower than the values of their neighbors.



Map of Local Moran classifications in 1994

Look for patterns and possible explanations for the clusters and spatial outliers in lung cancer mortality among white males. Also look for overall temporal trends as well as how locations change with time. The Local Moran method works best with data that is normally distributed. In some cases it is appropriate to transform the data before running the Local Moran method for this purpose.

NEXT STEPS

Outlier detection is a useful step in understanding patterns in data. It can often lead to questions that focus around trying to understand why outliers are different. You might look for other explanatory variables or testable hypotheses to explain why outliers exist. Another interesting avenue is to examine the dynamics of outliers in time or across spatial scales. They may be an artifact or anomaly at one particular time or scale, but persistence suggests that the outliers may be truly different from the other locations.