CC Spatial Cluster Training

Welcome to the spatial analysis portion of the Cancer Cluster presentation.

I am Frank Curriero, and in this training we'll be covering the basics for analyzing a spatial cluster.

We will cover several topics in this training, starting with an important definition to lay the groundwork.

For this training we will discuss the differences between clustering and cluster detection.

We will go over the different software and data used in this training and cover some examples so you can see a preview of what a real-life spatial cluster analysis might look like.

And finally we will conclude the presentation and give you some parting considerations and thoughts.

An important definition is that of spatial autocorrelation, which is the degree of similarity of nearby features.

In the context of cancer cluster analysis, this would mean how similar the number of cancer cases or cancer rates are in each geographic unit.

Spatial order correlation can either be positive or negative, with values between -1 and one.

If the autocorrelation is close to one, this would mean that the area is very tightly clustered with all geographic units being near each other having very similar cancer rates.

In the image below, let's assume that the shape colors represent cancer rates with similar cancer rates demarcated by the same color.

As you can see on the right side of the image, all the light green, dark green, and blue units are proximal to each other representing near similar units, and we would expect a high order correlation.

An order correlation of -1 would mean all units are perfectly distributed with no similar features near each other like shown in the left side of the image where none of the colors are touching.

An important distinction when performing an analysis would be to understand the differences between clustering and cluster detection and what they mean for your analysis.

For census tract level cancer rates, clustering would mean tracks that are closer together or adjacent have cancer rates that are more similar than tracks that are not close together or not adjacent.

Spatial order correlation techniques are used to assess clustering, such as Moran's I that can assess clustering over the entire map, and local versions such as local Moran's I to assess where on the map clustering is stronger or weaker.

The local versions are often referred to as local indicators of spatial association, or LISA.

LISA Cluster detection is defined as a geographically bounded sub area on your map, such as a collection of adjacent tracts whose combined cancer rate is anomalous compared to what is going on elsewhere on the map.

Anomalous can be on the high end like a hotspot, or the low end, like a cold spot.

Similar descriptions hold when dealing with spatial point level data such as address level data.

For this training, we will be running through a couple examples outlining techniques described in the guidelines.

Visualizing data is an essential part of understanding your data, and we will guide you through one way to visualize your cancer data using qualitative methods.

First, we will be looking at whether clustering exists using quantitative methods, showing the use of local Moran's I to visually check for high and low areas of clustering.

Last, we will show an example of cluster detection using the software SAT Scan to find statistically significant cancer clusters.

We're going to show a few examples of these techniques.

For these examples, we will be using two different softwares that are freely available for download, Geoda and SAT Scan.

Geoda is an open source tool with great documentation that will allow you to perform a large variety of qualitative and quantitative analysis.

SAT Scan is a cluster detection software that can analyze spatial, temporal, and space-time data using the scan statistic methodology.

While these are the tools being covered in this presentation, there are likely similar tools in your preferred platform such as Python or R.

The data sets used for this study come from the data that is available on the CDC Environmental Public Health Tracking Portal.

We will use the aggregated data to show examples of these tools.

Pennsylvania prostate cancer rates and standardized incidence ratios.

SIRs are available at the census tract level for a 10-year aggregation.

2010 to 2019 Georgia lung cancer rates and SIRs are available at the 5000-population aggregation, which is part of the tracking network Sub- County aggregations.

A point to stress, all analysis results and interpretations are for the sole purpose of demonstrating the software and processes that can be involved in a cancer cluster analysis.

They should only be viewed in this context.

We will first be looking at whether clustering exists using qualitative and quantitative methods, showing the use of local Moran's I to visually check for high and low areas of clustering.

First, visualizing data is an essential part of understanding your data, and we will guide you through one way to visualize your cancer data using chorifleth maps.

Choropleth maps help you display differences in values across spatial units.

Choropleth maps are thematic maps that use color to indicate the value of individual geographic units.

For example, in a cancer analysis, you could map the cancer rates by census tract or measure the level of a covariate per census tract for comparisons.

For example, this choropleth map was created using Geoda.

Age adjusted prostate cancer rates by census tract in Pennsylvania were categorized using manual breaks for the categories, with the higher rates being displayed in darker colors.

This choropleth map showed lung cancer rates in Georgia by the 5000 population areas that the CDC Environmental Public Health Tracking Program created to increase population counts within small geographies.

If interested in learning more about that, the link is provided below.

It shows the differences in rates over five-year periods.

Although some of the years of data overlap, this map helps to show where rates may have increased or decreased over the 15 years.

GeoDa allows you to click and explore on multiple connected maps.

Here we have highlighted 1 geography that has moved from the third-rate level to the highest rate level that we have categorized manually.

It also allows you to zoom in and take a look at the neighboring areas.

Next, we will look at measures of spatial order correlation using Local Moran's I.

When performing a Local Moran's I, we are interested in not just whether cancer rates are high or low, but the relationship between the neighboring geographic areas.

GIS software will produce a map highlighting areas in Reds and Blues that are called high, high, low, low, low, high, and high low.

When we say high, high, this means that there are geographic areas that have high values next to other geographic areas with high values.

So if the area of interest has a high cancer rate, we should ask, are the surrounding areas high as well or low?

This will result in one of the four combinations stated along with local Moran's I.

We will be doing a global Moran's I.

While a local Moran's I is interested in individual geographic units.

The global Moran's I looks at the entire area to check for overall order correlation of the data.

The value will be between -1 and one, like a typical correlation value.

When starting this analysis in GeoDa, it is important to first identify a weighting matrix which identifies the nearest neighbors of each census tract or geography.

More can be found at the link provided below.

Once the local Moran's I is run, several maps are produced.

The one on the slide is a connectivity map.

It allows you to check which tracts are connected to each other.

On the top image, you can see the neighboring tracts highlighted and the connection lines in the image below.

More details about how to do this are on the GeoDa website.

This slide shows the results of the local Moran's I.

The map on the top is showing the high, high, low, low, etcetera spots that we mentioned previously.

The red areas are tracks that have high rates and have neighbors whose rates are also high.

The areas in blue are showing tracks that have low rates and also have neighbors with low rates.

The lower map with green colors is the statistical significance map.

For the top map.

Levels of significance, for example 0.05 or 0.01 are shown in darker greens.

It is important when looking at the significance map to note what level of significance you're interested in.

GeoDa offers some important guidance to consider because of the multiple comparisons where AP value of point O 5 May not be appropriate to use.

The graph on the right side of the slide shows a Moran scatter plot and the global spatial order correlation value of .268 at the top, indicating fairly low positive spatial order correlation.

The matrix table allows you to look at specific geographies that may be of interest.

Here we have highlighted some of the outline points in the high-high quadrant.

Similarly, these are some results in Georgia where we are showing the importance of looking at changes over time.

Here we can see some areas where the rates are high.

Neighboring geographies are also high, which point to some places to potentially examine the data further.

After running our quantitative and qualitative clustering analysis, we can move on to cluster detection techniques.

For this analysis, we'll be using SAT scan, which uses Kulldorf's spatial scan statistics.

SAT Scan helps to detect spatial or space-time disease clusters to see if they are statistically significant.

The test essentially assumes whether disease is randomly distributed over space, over time, or over space and time.

SAT Scan can also perform prospective disease surveillance and can be applied to a variety of data types.

If you are interested in a more thorough explanation, SAT Scan has great documentation that can guide you through the process.

We're using SAT Scan today because it is free, widely used, and has a lot of applications to other fields.

This slide shows output of running SAT scan for Pennsylvania prostate cancer at the census tract level.

Cases are aggregated over the 10 year.

2010 to 2019.

This analysis was run for prostate cancer rates and the data input included case totals per tract, population, male 18 plus population per tract, and tract centroid coordinates.

Also included is information for which tracts had their data suppressed due to small numbers, cases or population.

SAT Scan can take this into account.

As a note, these images are for demonstration purposes only.

The map shows 20 areas for further evaluation.

SATScan identified the rates within these areas as significantly higher than the rates outside the areas at AP value set of 0.05.

Depending on your desired output, SAT Scan provides options for different types of graphical and text-based results.

Satscan can provide HTML and KML files for quick and automatic mapping of results in Google Maps and Google Earth as shown here, or shapefile format so users can import results into a GIS for more flexible mapping.

The text-based output option from SATScan provides detailed information about each identified cluster.

For example, as highlighted here, information about the Pittsburgh area cluster includes total number of cases, total population, a measure of relative risk, and the cluster P value which you see is highly significant.

This automatic HTML slash KML mapping interface from SAT Scan allows the user to click on any cluster to see this information, all of which can be saved in text-based output files from SAT Scan.

The previous analysis only adjusted for population at risk, the male population 18 plus, and we can understand the rationale for this.

We expect more cases in areas where we have larger populations.

The same can be said for known risk factors.

For example, age is a risk factor for prostate cancer, and we would expect more cases in areas with older male populations.

So we can rerun the SAT scan analysis, adjusting for the geographic distribution of age and identify clusters that cannot be explained by the male age population.

To run this adjusted analysis and SATScan, you'll need the age category by case breakdowns for each tract as well as the population by age category.

SAT scan generates an expected count based on the cases age distribution.

This is the same calculation that goes into calculating the standardized incidence ratio or SIR, and if those expected counts are available outside of SATScan, they can be brought in to run the adjusted analysis as was the situation here.

After adjusting for the age distribution, SATScan identifies 6 significant clusters compared to the unadjusted analysis or the analysis that just adjust for the population.

The age adjusted analysis shows that the total number of clusters decreases from 20 to 6.

Some cluster areas decreased in size and some cluster areas increased in size.

You can dive deeper into these interpretations providing further insight into how the age distribution may have explained or concealed clusters of prostate cancer.

We ran a similar analysis for the Georgia lung cancer.

For these data, we had lung cancer outcome for the three time periods 2006 to 2010, two 1010 to 2014, and 2014 to 2018, as well as expected counts based on age adjusted SIRS.

Results from the age adjusted cluster detection analysis show similar numbers of clusters and in some consistent areas over the three-time period.

In summary, we've shown a couple of different methods for analyzing spatial clusters, but these are not comprehensive.

The important thing to note is that each of these steps build upon each other and can highlight various patterns within your data.

The qualitative analysis gave us our initial look into the data and got us thinking about where higher rates may be occurring.

We also identified areas where higher rates were near other areas with high rates.

Finally, in our cluster detection analysis, we discovered statistically significant clusters where the observed values were higher than expected.

For both cluster detection and clustering analysis.

It's important to note that the data analytic methods, such as those described in the presentation, are just one component of an assessment.

For cluster detection, it's important to consider statistical significance along with practical significance.

Identifying spatial clusters is more of a beginning in cancer cluster analysis than an end.

This is particularly important as random clustering is always possible and not due to any particular cause.

Understanding the population demographics, socio demographics and environmental characteristics, for example of identified clusters can be important, as well as a multitude of other factors such as residential history, cancer site and cancer latency.

The methods in SAT scan allow for adjusted cluster detection analysis, that is to search for clusters that cannot be explained by other factors like known risk factors.

As was demonstrated, such factors are not only restricted to those that are socio demographic but can include for example, environmental factors for clustering.

Similar comments apply with interpretation to look into areas identified as being significantly similar.

Two final things to add.

If this information will be shared to the public, it needs to be placed in the proper context and should include a health communicator in discussions about the best way to relay the information.

As a final consideration, the clustering analysis conducted in Geoda and the cluster detection analysis performed in SATScan were for demonstration purposes only.

These software packages have great supporting resources with web-based material, tutorials and manuals.

You are encouraged to utilize these resources when performing your own analysis.

Thank you for your time and we hope this presentation clarified the basics and the start of the spatial portion of a cancer cluster analysis.