Spatial Clustering and Spatial Cluster Detection:

A Training Module for the CDC/ATSDR Guidelines for Examining Unusual Patterns of Cancer and Environmental Concerns

1. Introduction

As a companion to support implementation of the CDC/ATSDR Guidelines for Examining Unusual Patterns of Cancer and Environmental Concerns (hereafter referred to as 2022 Guidelines), the Geospatial Research, Analysis and Service Program is providing spatial clustering and spatial cluster detection trainings using both the GeoDa and SaTScan™ software. This document provides background information on the software and the data used. Separate documents provide step-by-step tutorials for using each software for the respective analysis, complete with screenshot displays, analysis results and summary interpretations.

This background document is organized as follows.

1. **Introduction** – introductory comments to orient the user
2. **What are Spatial Clustering and Spatial Cluster Detection Methods and Why Perform these Analyses?** – a summary of the difference between spatial clustering and spatial cluster detection and how these analyses fit into the 2022 Guidelines
3. **GeoDa Software** – instruction on how to download GeoDa and relevant software details
4. **SaTScan Software** – instruction on how to download SaTScan and relevant software details
5. **Data** – information on the CDC National Environmental Public Health Tracking Network data used in our tutorials
6. **Demonstrations using the Pennsylvania Prostate Cancer Data** – separate documentation providing comprehensive tutorials for using GeoDa for spatial clustering analysis and SaTScan for spatial cluster detection analysis
7. **What's Next after Evaluating Spatial Clustering and Spatial Cluster Detection Methods?** – list of potential next steps to align with the 2022 Guidelines

After completing these tutorials, users will be able to distinguish between spatial clustering and spatial cluster detection, use GeoDa for exploratory spatial data analyses (spatial clustering analyses), and use SaTScan for spatial cluster detection analyses.

2. What are Spatial Clustering and Spatial Cluster Detection Methods and Why Perform these Analyses?

As described in Appendix B of the 2022 Guidelines, Geographic Information System (GIS) can be useful for both mapping visualization and spatial analyses to understand geographic patterns at varying levels of geographies. These tools can be used during different phases of an examination and may help to highlight geographic areas that can benefit from further investigation. The guidelines suggest different methods for consideration depending on the questions that the investigator poses.

It is important to understand the differences between spatial clustering and spatial cluster detection and what they mean for your analysis. Although these terms appear to be similar, they have distinct meanings leading to different interpretations when analyzing geographic patterns. One of the tutorials provided focuses on using a software called GeoDa to assess spatial clustering. Spatial clustering examines spatial patterns and evaluates whether similar values or observations are close in proximity to each other. For evaluating cancer patterns, this might mean looking at rates in one geography (like a census tract) compared to neighboring census tracts. At the census tract level, clustering indicates that adjacent or nearby tracts have cancer rates more similar to one another than to tracts farther away.

Another tutorial is focused on using SaTScan for spatial cluster detection, which looks to identify areas on the map where observations are unusually high/low. For evaluating cancer patterns, this would involve identifying areas (e.g. a collection of adjacent census tracts) whose combined cancer rate is significantly higher than expected compared to the cancer rates elsewhere on the map.

3. GeoDa Software

GeoDa is a free software package that conducts spatial data analysis, geovisualization, spatial autocorrelation, and spatial modeling. The program provides a user-friendly, graphical interface for exploratory spatial data analysis (ESDA), such as spatial autocorrelation statistics for aggregate data and basic spatial regression analysis for point and polygon data. It was first developed by Luc Anselin and his team from the University of Illinois at Urbana-Champaign. GeoDa is continuously updated. Download the latest version from GitHub for either the Windows, MAC, or Linux platforms at https://geodacenter.github.io/.

The current GeoDa website https://spatial.uchicago.edu/geoda contains multiple trainings and practice datasets to help advance a user's understanding of the software. The cancer spatial cluster training materials in this document utilized the GeoDa tutorials. Several of the screenshot demonstrations are similar to those tutorials to maintain consistency using relevant cancer data for this specific tutorial.

In addition to producing maps, GeoDa can also produce histograms, box plots, and scatter plots to conduct exploratory data analyses.

4. SaTScan Software

SaTScan is a free software for the detection of spatial, temporal, and space-time clusters using scan statistics methodology. It offers flexibility in accommodating various data types with options for multiple types of analyses (e.g. space, time or space-time analysis, prospective based analysis and analysis adjusted for demographic information). To download SaTScan visit the website www.satscan.org and select "Download" in the lefthand margin and follow the instructions. To obtain the download password, you will need to register, providing your name, email address, affiliation and country.

The SaTScan website www.satscan.org provides resources for both new and experienced users. These resources can be found in the lefthand margin of the main landing page under the Download section. A few of these resources worth noting include the SaTScan technical documentation – a 100+ page software manual; a bibliography of published literature on use of the SaTScan software across multiple scientific disciplines; and a set of SaTScan tutorials with data that walk users through some of the different types of spatial cluster detection examples.

Regarding the latter, we would like to acknowledge *SaTScan Tutorial #1: Purely Spatial Poisson Scan Statistic for Cancer Incidence* (by Thomas Talbot, Sanjaya Kumar, Martin Kulldorff) which is based on New York cancer data. Our screenshot walk-through demonstration of a comparable analysis using cancer data from CDC (document provided separately), as well as information in some of these introductory sections, is modeled directly from this tutorial

SaTScan takes a model-based approach towards spatial cluster detection, meaning it uses a specified statistical model based upon the distribution of the data. For example, the Poisson model is used when data are aggregated to a geography representing a rate and both the numerator (number of cases) and denominator (population at risk) are available for analysis. This will be the type of analysis demonstrated in the separate SaTScan tutorial with the Pennsylvania prostate cancer data. Other types of common distribution models are Bernoulli and Normal. The Bernoulli distribution is used when locations represent a binary outcome, such as with cases and controls of a disease. The Normal distribution is used for continuous data such as with environmental exposure data. Other available models and technical details can be found in the SaTScan User Guide.

A final note, the SaTScan software can directly visualize analysis results using Google Earth. For this feature to work, you'll need to have Google Earth installed on your computer, which is free and available at http://www.google.com/earth/index.html. This is an optional feature.

5. Data

The data used for the demonstration is available on the CDC National Environmental Public Health Tracking Network Website or at the specific links provided below. Pennsylvania prostate cancer rates (males 18+ years old) and standardized incidence ratios (SIRs) are available at the census tract level for a 10-year aggregation (2010-2019). SIRs for Pennsylvania prostate cancer are census tract age-adjusted rates per 100,000 population. If you are interested in learning more about these data, please visit the link below.

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

The Pennsylvania data are available for these trainings in shapefile and Excel file formats. Variables were added to these files specific for these trainings which included the latitude and longitude coordinates for each tract's geometric centroid as well the age-adjusted expected prostate cancer count for each tract (the denominator of the SIR). The Excel file was also pared down to just include the variables needed for the trainings.

Details about specific variables are described in the tutorial documents. In addition, a complete data dictionary for the Pennsylvania data is provided along with these training documents. Please note that some data are missing due to data suppression rules, and for ease in these trainings we have generated a separate csv file of the missing/suppressed data. This file will be identified and described in the training materials.

6. Demonstrations using the Pennsylvania Prostate Cancer Data

In separate documentation we provide step-by-step tutorials for using GeoDa for spatial clustering analysis and SaTScan for spatial cluster detection analysis.

7. What's Next after Evaluating Spatial Clustering and Spatial Cluster Detection Methods?

As mentioned before, spatial clustering and spatial cluster detection methods are tools that can be used in a more comprehensive analysis. The purpose of these trainings is to share some methods that might be used to help evaluate spatial clusters in an examination of unusual patterns of cancer.

The methods shown in these tutorials can be used along with the 2022 Guidelines Decision Making Form, provided as a resource/tool at (https://www.cdc.gov/cancer-environment/media/pdfs/Decision-Making-Form-508.pdf), to get a better understanding of both cancer and environmental factors in areas of concern. Advanced analyses beyond these trainings include but are not limited to:

- Space-time cluster detection and other options in SaTScan
- Adjustment of analyses for additional demographic variables
- Adjusting analyses based on factors outside of SaTScan (behavioral, environmental, etc.)

If your analysis did not find clusters present, or after adjustments clusters are no longer statistically significant, it does not necessarily mean that analyses are complete. Investigators might consider other potential risk factors or trends outside the spatial landscape. For example, cancer characteristics such as site, latency, and residential history are additional factors that can impact cancer investigations. In addition, if your analysis did find spatial clusters 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. It is important to consider some limitations with these analyses including: 1) spatial clusters do not imply causality, 2) suppressed or missing data may bias results, and 3) small-area analyses are sensitive to population denominators.

In summary, spatial clustering involves assessing correlation or areas on the map where cancer rates appear similar with their surrounding neighbor rates and spatial cluster detection identifies areas on the map where cancer rates are significantly higher or lower than expected. Spatial clustering and spatial cluster detection are just two methods for examining unusual geographic patterns. Other methods that might help to further characterize spatial patterns, such as statistical regression methods for spatial data, offer opportunities to quantify determinants (e.g. clinical, demographic, environmental). Further discussion of additional methods is provided in Appendix B of the Guidelines for Examining Unusual Patterns of Cancer and Environmental Concerns. As emphasized in the 2022 Guidelines, these methods should be considered as screening tools or early steps in the broader examination process. They should be integrated with epidemiological reasoning, decision-making frameworks (e.g., the CDC Decision-Making Form), and consideration of non-spatial risk factors.

References / Further Reading

1. CDC/ATSDR Guidelines for Examining Unusual Patterns of Cancer and Environmental Concerns, https://www.cdc.gov/cancer-environment/php/guidelines/index.html
2. CDC National Environmental Public Health Tracking Network, https://ephtracking.cdc.gov/DataExplorer/
3. GeoDa Software: https://geodacenter.github.io/, https://spatial.uchicago.edu/geoda
4. SaTScan Software: http://www.satscan.org
5. Anselin L. Local indicators of spatial association—LISA. Geogr Anal. 1995;27(2):93-115.

6. Kulldorff M, Nagarwalla N. Spatial Disease Clusters: Detection and Inference. Vol 14; 1995.
7. Kulldorff M. A spatial scan statistic. Commun Stat methods. 1997;26(6):1481-1496.
8. Waller LA, Gotway CA. Applied Spatial Statistics for Public Health Data. New York: John Wiley and Sons; 2004.
9. Huang L, Pickle LW, Das B. Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. Stat Med. 2008;27(25):5111-5142. doi:10.1002/sim.3342.