ClusterSeer®

software for the detection and analysis of event clusters

User Manual
book 2
version 2.5

BioMedware
Geospatial Research and Software

Project Leaders: Geoff Jacquez and Leah Estberg

STTR Collaborating Institutions: BioMedware, Inc., the University of Michigan, and the University of Minnesota.

Software developers: Leah Estberg, Andrew Long, Eve Do, and Bob Rommel.

Manual and help authors: Heidi Durbeck, Dunrie Greiling, Leah Estberg, Andrew Long, Geoff Jacquez, Yanna Pallicaris, and Susan Hinton.

Advisors: Luc Anselin, Arthur Getis, Dan Griffith, Uriel Kitron, Lance Waller, and Mark Wilson.

The following individuals provided suggestions and insights that greatly improved the software: Peter Diggle, Richard Hoskins, Martin Kulldorff, Bruce Levin, Tonny Oyana, Peter Rogerson, and graduate students and instructors in the course "Spatial Epidemiology" offered in 1999 & 2000 at the School of Public Health, University of Michigan.

This project was supported by STTR grant #CA64979 from the National Cancer Institute to BioMedware, Inc. The software and manual contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute.

Cartographic boundary files of ZIP codes in Nassau, Suffolk and Queens counties, New York were provided by Claritas (Source: Claritas Inc./Geographic Data Technology, Inc., 2001).

ClusterSeer is protected by U.S. patents 6,360,184, and 6,460,011.

For updated troubleshooting information and Help please visit ClusterSeer online (www.biomedware.com/?module=Page&sID=clusterseer-help-and-tutorials).

Preface

ClusterSeer provides data visualization tools and state-of-the-art statistical methods to explore spatial and temporal patterns of disease and other events.

ClusterSeer methods can be used to investigate clusters of events in space, in time, and spatial clusters that depend on time (spatio-temporal interaction).

Use the method of your choice, or find an appropriate method using the ClusterSeer Advisor available within the software.

System requirements

- Windows 95 or Windows NT 4.0 or more recent operating system
- Screen resolution of 800 X 600 or finer for best viewing of the maps and graphics
- 256 colors or better highly recommended for graphics
- 15 MB hard drive
- 128 MB RAM (minimum), 256 MB RAM (recommended)
- P3 500 MHz (recommended) Processor Speed

Manual overview

This manual outlines how to use ClusterSeer, BioMedware's tool for detecting pattern in event data.

The first ClusterSeer manual presents the conceptual background for the software, provides an overview of how to use ClusterSeer, and describes 10 cluster detection methods. The second ClusterSeer manual is a continuation and expansion of the first. Chapter 16-What's New in ClusterSeer 2- outlines the new methods, concepts and features in the software. Chapters 17-31 describe the 15 new statistical methods.

The manual also has a resources section that includes a glossary, references, and an index.

For easier differentiation of interface and description, this manual will use the following style conventions:

Typeface	Meaning
serif type	explanatory text
sans serif type	part of the ClusterSeer interface, such as menu items or dialogs

This information is also available in online help accessible from the the "Help" menu and the "Help" buttons on dialogs in ClusterSeer. The online help has hyperlinks that connect related topics.

BioMedware also has ClusterSeer Online Help on its website, www.biomedware.com/?module=Page&sID=clusterseer-help-and-tutorials. Please check this for updates and additional information.

Contents

CHAPTER 16

What's New in ClusterSeer 2

ClusterSeer provides data visualization tools and state-of-the-art statistical methods to explore spatial and temporal patterns of disease and other events.

ClusterSeer now offers 15 new methods to investigate event clusters in space, in time, or spatial clusters that depend on time (spatio-temporal interaction). ClusterSeer also includes 4 new statistical concepts and 8 new features. This chapter details these new methods, concepts and features.

New methods

New Methods	Spatial Clustering		Temporal Clustering		Spatio-Temporal Clustering
	Global	Local	Single	Multiple	
Cuzick & Edwards' Method	■				
Dat's Method			■	■	
Direction Method					■
Ederer-Myers-Mantel Method				■	
Empty Cells Method			■	■	
Getis-Ord Local *G* Method		■			
Grimson's Method	■		■	■	■
Jacquez's *k* Nearest Neighbor Method					■
Knox Method					■
Kulldorff's Scan Method		■			
Larsen's Method			■	■	
Mantel's Method					■
Moran's *I* Method	■				
Oden's I (Pop) Method	■				
Scan Method			■	■	

New concepts

- Nearest in space
- Nearest in time
- Randomization types
- Statistical distance test statistic

New features

- Save project sessions
- Export maps, histograms, and plots as bitmaps
- Export histograms and plot data as DBF files
- Export mapped results as a shapefile
- Load in spatial features
- Restart session
- Legend pane for maps
- Shapefile requirements
- Import DBF files
- Import temporal data formats more easily

Methods for Detecting Global Spatial Clusters

Global cluster detection methods are used to investigate the presence of spatial patterns anywhere within the study area. They attempt to answer the question: Are there any unusual spatial patterns? These focus on whether clustering exists or not, regardless of location or scope. Essentially, these methods evaluate whether a spatial pattern exists in the data that is unlikely to have arisen by chance. The null hypothesis for these methods is simply "no clustering exists."

Individual-Level Data	**Group-Level Data**
Cuzick & Edwards	Besag and Newell's Method
Ripley's *K*-function	Moran's *I*
	Oden's *Ipop*
Grimson's Method	

Grimson's Method can be used with either individual or group level data. For surveillance of spatial data, use Rogerson's Method.

Methods for Detecting Local Spatial Clusters

These cluster detection methods are used to investigate spatial clusters in a particular area. They can be thought of as methods that attempt to answer the question: Are cases neighboring a particular case closer together than expected by chance?

Local cluster detection methods are available for group-level data only.

- Besag and Newell's Method
- Turnbull's Method
- Anselin's Local Moran
- Getis-Ord Local *G* Test
- Kulldorff's Spatial Scan Test

For surveillance of spatial data, use Rogerson's Method.

Methods for Detecting Temporal Clusters

Temporal cluster detection methods are used to investigate whether events (such as cases of disease) tend to aggregate in particular time periods. All are used on group-level data, though Grimson's method can be used on individual-level data as well. These methods can be used to evaluate disease frequency or case counts in a single or in multiple time series.

Method	Disease or event frequency	Case or event count, single time series	Case or event count, multiple time series	Case or event & population -at-risk counts
Dat's Method		■	■	
Ederer-Myers-Mantel Method			■	
Empty Cells Method		■	■	
Grimson's Method	■	■	■	
Larsen's Method		■	■	
Levin & Kline's Modified CuSum				■
Wallenstein's Scan		■	■	

Grimson's method will analyze spatio-temporal data, similar to multiple time series data.

Methods for Detecting Space-Time Clusters

Spatio-Temporal methods detect event clusters in space that depend on the time period (Space-Time interaction).

Individual-Level data	**Group-level data**
Direction Method	Kulldorff's Spatio-Temporal Scan
Jacquez's *k*-Nearest Neighbor Method	Grimson's Method
Knox's Method	
Mantel's Method	
Grimson's Method	

Methods that Adjust for Population

Several ClusterSeer methods adjust for population, because the population size can influence the likelihood of events such as disease transmission. There are three subgroups of methods that account for population in different ways. The first group of methods requires the user to submit disease or event frequency data. The second group requires case or event count and population-at-risk data to be submitted in separate columns. The third group, a case-control group, requires users to judiciously select controls from the same population as the cases. The table below shows which methods fall into each of these population adjustment categories.

Method	**Means of Adjusting for Population**		
	Disease or event Frequency	Case or event and population at risk count	Controls must represent the population at risk
Anselin's Local Moran Method	■		
Besag & Newell's Method		■	
Bithell's Method		■	
Cuzick & Edwards' Method			■
Diggle's Method			■
Getis-Ord Local *G* Method	■		
Kulldorff's Scan Method		■	
Levin & Kline's Modified CuSum		■	
Moran's *I* Method	■		
Oden's *IPop* Method		■	
Rogerson's Method		■	
Score Test		■	
Turnbull's Method		■	

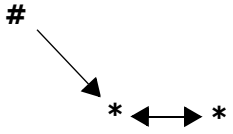
Concepts: Nearest in Space

Nearest neighbor relationships are part of methods such as Cuzick and Edwards' and Jacquez's *k*-Nearest Neighbor (*k-NN*) methods. These methods consider whether events neighbor each other in space or in space and time, respectively. Considering nearest neighbors avoids the problem of setting a threshold distance to evaluate whether cases are near or far from each other. Threshold distances are used in other methods, but may not be appropriate to all datasets. For example, if your dataset consists of both urban and rural locations, distances to neighbors will be longer in the rural locations than in the urban region. Thus, no single threshold distance will capture the types of neighborhoods you wish to consider.

Spatial nearest neighbors

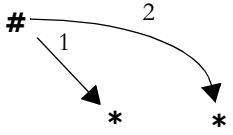
ClusterSeer calculates nearest neighbor relationships from the distance between events submitted as a file of point locations (as occurs in the Cuzick & Edwards' method).

Nearest neighbor relationships in space may or may not be reciprocal. A point could be its nearest neighbor's nearest neighbor, but perhaps its nearest neighbor is actually closer to something else. In the figure below, the asterisks are nearest to each other but the #'s nearest neighbor is a 1.



k-NN

A point has a nearest neighbor, but nearest neighbor relationships can be considered at higher levels. The nearest neighbor methods in ClusterSeer are flexible and can consider several levels of neighbors (first nearest neighbor, first and second nearest neighbor, etc.). *k* defines the number of neighbors to consider in the analysis. In the illustration below, the # has both asterisks as its first and second nearest neighbors (*k* = 2).



Ties

A problem with nearest neighbor methods is how to resolve ties. If two neighbors are the same distance from the event considered, which one should be scored? ClusterSeer solves the tie arbitrarily by choosing only one of the tied events.

Concepts: Nearest in Time

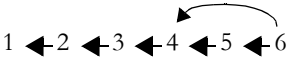
Nearest neighbor in time relationships are part Jacquez's *k*-Nearest Neighbor (*k-NN*) method.

Temporal nearest neighbors

Jacquez's *k-NN* method considers temporal adjacency to evaluate clustering. It categorizes whether events neighbor each other in space and time. ClusterSeer calculates nearest neighbor relationships in time by placing the events in a dataset in chronological order. The nearest neighbor of an event is the prior event, as illustrated in the diagram below. Because the nearest neighbor is always the prior event, nearest neighbor relationships in time are never reciprocal, unlike spatial nearest neighbors.

k-NN

A point has a nearest neighbor, but nearest neighbor relationships can be considered at higher levels. The nearest neighbor methods in ClusterSeer are flexible and can consider several levels of neighbors (first nearest neighbor, first and second nearest neighbor, etc.). *k* defines the number of neighbors to consider in the analysis. In the illustration below, 4 & 5 are the two nearest neighbors of 6.



Ties

A problem with nearest neighbor methods is how to resolve ties. If two neighbors are the same time from the event considered, which one should be scored? A problem with nearest neighbor methods is how to resolve ties. If two neighbors are the same distance from the event considered, which one should be scored? ClusterSeer solves the tie arbitrarily by choosing only one of the tied events.

Concepts: Types of Spatial Randomization

Monte Carlo randomization is one way to quantitatively evaluate observed data and test statistics. See the first ClusterSeer manual (p.20) for details on procedures for calculating Monte Carlo P-values. Within ClusterSeer, spatial randomization techniques vary among methods. For the multinomial and Poisson distributions, ClusterSeer generates random values by choosing values from the specified distribution. For conditional randomness, data values are reassigned among sub-groups.

Randomization Technique	**Cluster Detection Method**
Drawing from a multinomial distribution	• Besag and Newell's • Bithell-conditional • Kulldorff's Scan • Oden's *Ipop* • Turnbull's
Drawing from a Poisson distribution	• Bithell-unconditional • Score
Conditional randomness	• Local Moran
Randomize data (rate, count, or case-control label) among spatial locations by swapping labels	• Cuzick and Edwards' method • Getis-Ord Local *G* • Moran's *I*
Alter distances between points by multiplying their locations by a random number	• Ripley's *K* function

Concepts: Types of Temporal Randomization

Within ClusterSeer, temporal randomization techniques and distribution theory vary among methods. For Poisson distributions, ClusterSeer generates random values by choosing values from the specified distribution.

Randomization or Distribution Technique	Cluster Detection Method
These methods assume the cases are allocated with equal probability across the time cells. They each take *n* cases and distribute them at random among *t* time cell intervals. Unlike multinomial randomization techniques, these temporal methods have no population base information: thus, the probability of a case being placed in a particular time interval is not proportional to the population-at-risk size in that interval.	• Dat's method • Ederer-Myers-Mantel • Empty Cells • Larsen's method • Scan method
Drawing from a Poisson distribution	• CuSum • Grimson's method
Drawing from the Normal distribution	• Grimson's method

Concepts: Types of Space-Time Randomization

Randomization Technique	Cluster Detection Method
Shuffling time distances or adjacencies	• Jacquez's *k-NN* • Knox • Mantel
Shuffling time of occurrence of cases across case locations	• Direction

For the Knox, *K-NN*, and Mantel methods, ClusterSeer randomizes the space-time relationships by shuffling the time distances between cases or events while holding the spatial distances constant. The statistics are then recalculated on the randomized datasets. The null hypothesis for each of these tests is that there is no significant relationship between the spatial and temporal distances, so that breaking them should be no problem. If there is significant space-time association in the dataset, the random shuffling of the times will tend to produce datasets with less space-time association, and the observed value will be significant when compared to the randomizations.

For the Direction method, the significance of the average direction is evaluated through a randomization procedure which holds the sine and cosine matrices constant and randomly assigns connections between pairs of cases. This is equivalent to holding the locations of the cases fixed while randomizing their times of occurrence. This randomization procedure is repeated to generate a distribution of the angular concentration under the null hypothesis. A P-value is determined by comparing the angular concentration from the original (not randomized) data to this null distribution.

Concepts: Statistical Distance Test Statistic

This statistic is used to evaluate the significance of multiple sets of Monte Carlo simulations in Jacquez's *k*-Nearest Neighbor and Cuzick & Edwards' methods.

It combines the P-values across the number of tests you specify (*k*). Similar to the Bonferroni and Simes combined P-values, this statistic gives an overall probability that accounts for multiple comparisons. With this measurement, you can calculate the distance between the mean of a cluster J_{i1} and a single data point, *i*.

Allow *J* to signify a 1 X 10 vector of the test statistics ($J_1,\ldots J_{10}$) (You can substitute T_k for *J*). For each randomization ClusterSeer computes a *J* vector, which can be represented as a location in 10 dimensions. The results under randomization form a cloud of "Number of runs" points in this *10-D* space. The center of the cloud is the centroid. You can evaluate significance by comparing the statistical distance from the centroid of the observed vector *J* to the statistical distances from the centroid of the *J* vectors being randomized. The statistical distance from each point to the centroid is as follows:

$$d_i = \sqrt{\frac{(J_{i1} - \overline{J_{i1}})}{s_1} + \ldots + \frac{(J_{i10} - \overline{J_{i10}})}{s_{10}}}$$

Here d_i is the distance from point *i* to the centroid. J_{i1} is the value of the statistic, and J_k calculated for k=1 using the data from the first randomization. s_1 signifies the standard deviation of the J_1 under randomization, and $\bar{J}_1$ is the mean of J_1.

ClusterSeer calculates an upper-tail P-value based on the Monte Carlo simulations, counting the number of distances to the centroid that are greater than or equal to the distance from the observed *J* to the centroid. This P-value is the probability, under the null hypothesis, of observing a vector of J_k or (ΔJ_k) as or more extreme than the observed. If the combined P-value is smaller than 0.05, you can reject the null hypothesis that there is no spatial clustering.

New Features

Save your project session

You may save your work as a project to reopen in a later ClusterSeer session. The project file (*.csr) includes the session log, and any corresponding maps and plots you created, as long as the plots were generated without using Monte Carlo Randomization techniques.

Note: ClusterSeer does not save the histograms or plots created from Monte Carlo runs. However, you can export them to a bitmap or DBF format and view them that way. You can also run the analysis again to see the new histogram and plot, but because of the randomizations the outcomes may be slightly different from the original.

Export images

After performing a statistical analysis, you may export plots, maps, and histograms as images. to a bitmap, DBF, or Shapefile format. ClusterSeer allows you to export only those items that each statistical method includes in its output. The session log can be exported as a text file only.

1. Choose Export from the File menu.
2. Select the item you want to export (histogram, map, or plot) and choose the Bitmap format.
3. If you wish, you can change the default name of the file and where it will be saved. Then hit Save as. You can then open this image in any software program that accepts bitmap format files (*.bmp). You can also use an image processing program that can convert bitmaps to different formats.

Export histogram and plot data

You can also export statistical results as DBF files that can be imported into spreadsheet programs for further graphical refinement or statistical exploration.

1. Choose Export from the File menu.
2. Select the item you want to export (histogram or plot) and choose DBF format.
3. If you wish, you can change the default name of the file and where it will be saved. Then hit Save as.

You can then open this file in any software program that accepts DBF files (*.dbf).

Export mapped results as a shapefile

You can also export mapped results as shapefile that can be imported into a GIS program to layer with other data. ClusterSeer does not re-export data you brought in to the system, only data that result from analyses in the software.

See the table below to learn what types of information can be exported from which of the methods. For those methods that produce no mapped results (such as many global clustering methods) the only way to export the map is as an image.

1. Choose Export from the File menu.
2. Select map and choose shapefile format
3. Specify the coordinate system you want to use. This option is only available if you brought in geographic data (latitude-longitude decimal degrees). Otherwise, the results will be saved out in the same planar projection of your original dataset.
4. If you wish, you can change the default name of the file and where it will be saved. You must use a different name than your original data if you do not intend to overwrite it. Then hit Save as.

You can then open these file in any software program that accepts shapefiles.

Mapped results	Methods	Exported
Circular cluster outlines	Besag & Newell's Kulldorff's Scan Local *G* Turnbull's	Shapefile containing circular polygons that describe the cluster extent.
New variables for areas/ polygons	Local Moran	A new polygon shapefile with the single variable you chose to import and the new variables as added columns. Do not overwrite your original dataset, as you will lose any other columns from that file.
Space-time interaction lines between point locations	Direction Knox's	Shapefile containing lines that connect point locations (one set of lines for Direction, three types for Knox)
No mapped results (map only shows original data)	All other spatial, space-time, temporal, and surveillance methods.	No data exported. You can export a bitmap image of the mapped data.

Load in spatial features

Spatial features are vector files that contain locations or spatial information to help visualize spatial data and results. ClusterSeer can accept points as text or shapefiles or polygon shapefiles for use as spatial features. No data aside from the spatial boundaries or locations is imported.

For instance, if you are analyzing disease rates assigned to county centroids, you may wish to bring in a feature of the county boundaries so that you audience can more easily recognize the area's geography.

To load in a spatial feature, choose Import Spatial Feature from the View menu.

Restart session

If you wish to start your ClusterSeer session from scratch, you can choose Restart session from the File menu.

This clears the Session Log of the summaries of actions previously performed by ClusterSeer and any notes you have added. It also closes any open maps, plots, or histograms from the previous analysis.

Legend pane for maps

Most ClusterSeer maps are displayed in a three-pane window. The left-hand pane lists the active layers in the map, the middle pane contains the map itself, and the right-hand pane is the map legend. This legend pane is new in ClusterSeer version 2.

The right panel identifies the symbols for active map layers. You may need to expand the frame to view the full legend names. Displayed legend items are often color-coded to match the layers.

Shapefile requirements

ClusterSeer 1 could take shapefiles for some of the methods. Now, it accepts shapefiles as the main data format for all spatial methods.

ClusterSeer will send you an error message if your data do not meet its shapefile requirements. Make sure to prepare your data with a GIS data editor so that it does not contain self-intersecting polygons. A polygon is called "self-intersecting" when two or more of its borders intersect anywhere except their endpoints.

If your polygons overlap, it may be difficult to view them when mapped or to select them for queries. ClusterSeer will not be able to display properly shaded areas where overlap occurs. Uniquely named polygons completely contained within another polygon will be correctly processed for analysis and display. Relatively smaller, non-uniquely named polygons will be discarded on import and excluded from the analysis.

Import DBF files

With ClusterSeer, you can now import DBF files for all methods except Moran's I, Local Moran, Oden's *Ipop*, and Grimson's.

- For spatial and spatio-temporal methods, when you import a DBF file, ClusterSeer will prompt you to select which columns in the data file hold the relevant information. You must include labels when importing DBF files into ClusterSeer.
- For temporal methods with a single or multiple time series, your first column must contain the label, and subsequent columns should contain your case count data. You must include labels when importing DBF files into ClusterSeer. Modified CuSum is currently the only method that does not require a label.

You may keep track of your columns by labeling them in the first row of your dataset. The labels are separate from the first column of ID labels, and will not interfere with ClusterSeer's analysis.

Temporal data formats

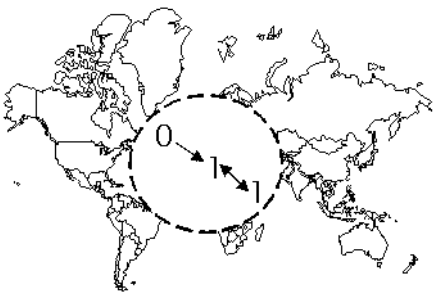
ClusterSeer 2 has relaxed its temporal data format requirements from what is described in the first ClusterSeer User Guide. Now there are two format options for all calendar-based time intervals (see format options 1 & 2 in the table below).

For the years, you can omit the first two numbers of the date for any date in the 1900s. Thus, if you use "89" for example, ClusterSeer will assume that the preceding numbers were "19" or "1989." You can use dates with two and four numbers in the same file as long as the dates other than those in the 1900s have four digits.

Sample data	Format Option 1 (with example)	Format Option 2 (with example)	Notes	Valid Range
Yearly	YYYY (1998)	YY (98)		01 to 9999
Monthly	YYYYMM (199801)	YYMM (9801)	monthly values (MM) range from 01-12	0101 to 999912
Weekly	YYYYWW (199843)	YYWW (9843)	weekly values (WW) range from 01-52	0101 to 999952
Daily	MM/DD/YYYY (1/2/1998)	MM/DD/YY (1/2/98)	month and date values may be expressed as single digits	12/30/1899 to 12/31/9999 or 12/30/00 to 12/31/99
User-defined	User-defined (5)	User-defined (5)	Positive whole numbers that may represent points in time or non-overlapping, successive temporal intervals. In this scale, the intervals are naturally ordered by their magnitude (5 comes after 4) and there is a known unit distance between any 2 successive numbers.	0 to 4.2 billion

CHAPTER 17

Cuzick & Edwards' Method



Cuzick & Edwards' method (Cuzick & Edwards 1990) can detect global spatial clusters in individual-level case-control data. This method uses the control location to reflect the geographic variation in the population density as a whole. Use this method when you know both case and control locations (e.g. place of residence), and when you have selected controls from the same population as the cases. You should code your cases 1 and your controls 0.

Examples

Dockerty et al.(1999) used Cuzick and Edwards' method to analyze clustering in leukemias and lymphomas among young people in New Zealand, a country without nuclear installations. They found that there was no statistically significant spatial clustering in any of the leukemias or lymphomas tested. Doherr et al.(1999) found significant spatial clustering of *Cornyebacterium pseudotuburculosis* in horses using Cuzick and Edwards' method. They used information on spatial clustering to infer transmission patterns of the infection.

Cuzick and Edwards' Method: Statistic

H_o	Cases & controls are sampled from a common spatial point distribution.
H_a	The cases are spatially clustered relative to the controls.

For this method, ClusterSeer quantifies nearest neighbor relationships between individuals to determine whether clustering exists in the dataset.

Test statistic

The test statistic, T_k, counts how many cases neighbor other cases. You can define a number of different sizes of "neighborhoods" or spatial scales by specifying k, which indicates the number of nearest neighbors to consider in the analysis. ClusterSeer calculates the number of k nearest neighbors to each case that are also cases. T_k is the total over all cases in the dataset. For example, T_1 is the number of cases in the dataset neighbors are also cases (k=1).

$$T_k = \sum_{i=1}^{N} \delta_i d_i^k$$

Where:

- N: the sample population size
- N_0: the number of cases
- N_1: the number of controls
- $\delta_i = 1$ if observation i is a case and 0 if it is a control.
- $d_i^k = 1$ if the kth nearest neighbor to i is a case, 0 otherwise.

The expected value of the test statistic under the null hypothesis is

$$E(T_k) = pkN$$

In this equation,

$$p = \frac{N_0}{N}\left(\frac{N_0 - 1}{N - 1}\right)$$

The z-score is calculated as:

$$z = \frac{T_k - E(T_k)}{\sqrt{Var(T_k)}}$$

The z-score calculates a standardized difference between the observed T_k and expected $E(T_k)$ values of a statistic divided by the standard deviation. For more information about z-scores, see the first ClusterSeer User Guide, p. 19.

Cuzick & Edwards' Method: Significance

ClusterSeer provides several ways to evaluate the significance of the test statistic (T_k).

- ClusterSeer provides the upper-tail P-value, which is the probability under the null hypothesis of observing a T_k as large or larger than the one given in T_k. It is based on the assumption of a normal distribution of the data.
- ClusterSeer also generates P-values via the Monte Carlo simulations for each k. ClusterSeer randomizes the data by shuffling the case-control labels for each of the spatial locations. This is a way to compare the observed T_k to the distribution of T_k based on a random distribution of the data.

Multiple comparisons analysis

When k>1, the method calculates a statistic at each value of k. This is multiple testing. ClusterSeer automatically runs a multiple comparisons analysis to determine the proper significance level for all comparisons. ClusterSeer provides a combined P-value for all tests performed at one initial alpha level. This is accomplished through Bonferroni and Simes adjustments.

- Bonferroni $P_c = j[min(P_i)]$
- Simes $P_c = min(n+1-i)P_i$

In this case, P_c denotes the combined P-value for all tests, P_i the value for an individual test, j is the number of comparisons, and i is the sequential index for the individual test considered. You can compare this value to your original alpha level to see if the tests show significant results.

This topic is related to the combined P-values feature available in multiple comparisons for other methods, but it uses Simes' formula instead of Holm's.

Cuzick and Edwards' Method: How To

Choose Cuzick & Edwards' method from the Analysis menu. (Spatial|Global submenus).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data file (text file, DBF, or point shapefile).

   If you submit a text file with header, DBF, or shapefile, ClusterSeer will prompt you to identify which columns in your file contain the required data. In this case, the columns can be in any order

   If it is a text file without a header, it should contain individual-level data with the following columns in the following order.

Subject label (optional)	Point x-coordinate (not for shapefile)	Point y-coordinate (not for shapefile)	Case/control status

2. You will need to specify the coordinate system of the data. If the data are in geographic coordinates, you will also need to choose a distance measurement.
3. If your data file includes labels, choose Selected data file contains label. If your dataset has no column of labels, select Use study row # as label.

Choose settings

4. In the Provide data dialog, you may use the Select File button to change your file choices.
5. Enter the maximum number of nearest neighbors to analyze for clustering. The size of the cluster you choose to detect *(k)* determines in part where you can detect significant clusters. The value you choose to enter should be based on information about the disease process.
6. Choose the number of Monte Carlo runs, the number of simulations used to determine statistical significance of the test statistic. The default value is 999.
7. Enter the significance level you wish to use for the test. The significance level is the alpha level, the cutoff for statistical significance. The default value is 0.05.

Run the analysis

8. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.
9. Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Cuzick & Edwards' Method: Results

Distribution

You can view the Monte Carlo distribution by choosing MC Distribution from the View menu.

The histogram shows the reference distribution generated by randomizing the dataset and recalculating T_k. Select k, the number of nearest neighbors you would like to see in the distribution. T_k is illustrated in red, and it is compared with the distribution for estimating the one-sided P-value.

Map

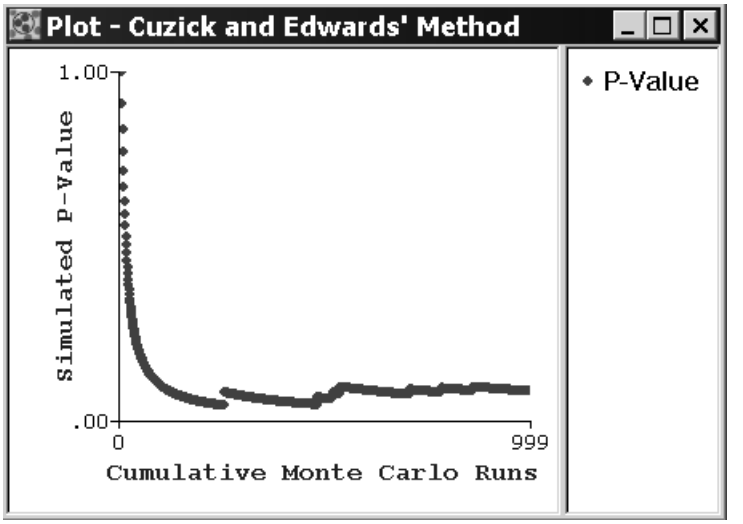
You can view the map by choosing Map from the View menu. The map has one layer that shows case-control points.

It can be queried to show the x- and y-coordinates of the point, the subject label and case-control status. You may use the zoom function to magnify the points.

Plot

Choose Plot from the View menu. You can view a plot of P-values simulated from Monte Carlo runs. Using the plot may help you to determine the minimum number of Monte Carlo runs to perform in your analysis. The simulated P-value plot shows how the significance of the test statistic changes with the number of Monte Carlo randomizations performed.

What you will usually see is that the P-value decreases from near p=1.0 to an asymptote before it reaches the number of randomizations you specified in the analysis. If it decreases to the asymptote after few randomizations, you specified a greater number of randomizations than was required to find the P-value. If it continues to fluctuate, you may wish to rerun the analysis with a greater number of randomizations to find the P-value.



Session log

After ClusterSeer performs a Cuzick and Edwards analysis, it will place summary information and results into the session log.

Parameters and summary statistics

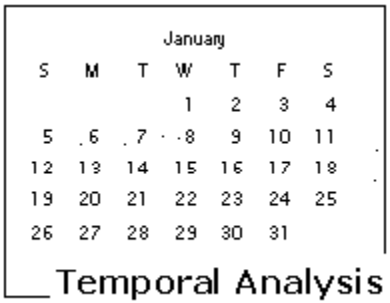
- k, the number of nearest neighbors
- The Test statistic, T_k, counts how many cases neighbor other cases.
- The expected value of the test statistic $E(T_k)$ under the null hypothesis
- The variance of T_k under the null is a fairly complex expression and is given in Cuzick and Edwards (1990). The variance describes the amount of variability of your data around the mean value.
- Z-scores
- Upper-tail P-value: the probability under the null hypothesis of observing a T_k as large or larger than the one given in T_k, based on a normal distribution of the data.
- Monte Carlo P-values

Combined P-values

- Bonferroni and Simes corrected P-values for the normal approximation, and for the Monte Carlo Distribution: See "Multiple comparisons analysis" on page 180.
- Statistical Distance Test statistic and its P-value: Refer to "Concepts: Statistical Distance Test Statistic" on page 169.

CHAPTER 18

Dat's Method



Dat's 0-1 Matrix Test (Dat 1982) can detect temporal clustering in single and several time series with group-level data. Use Dat's method with counts, not rates, on 5-10 time intervals. Within a time series the method assumes population size does not change through time.

You cannot use Dat's method when the expected number of cases in each interval is smaller than 2. This means the total number of cases in the time series must be greater than twice the number of time intervals. Otherwise, consider using the Empty Cells method. See "Empty Cells Method " on page 211. If 11 or more time intervals are present you may be able to aggregate data by combining time intervals.

Dat's method is more sensitive than the Ederer-Myers-Mantel method in detecting multiple clusters within a space sub-unit.

ClusterSeer's temporal methods do not analyze space-time interactions. If you have data pertaining to spatial location, you may want to consider running one of ClusterSeer's Spatio-temporal cluster detection methods. See "Methods for Detecting Space-Time Clusters" on page 161.

Dat's Method: Statistic

H_0	Cases occur at random over the *t* time periods.
H_a	Cases do not occur randomly through time.

ClusterSeer provides two tests for temporal clustering under Dat's method: within a single time series (using the z-score) and across several time series simultaneously (using the Chi-squared statistic).

The test statistic, *A*, is the number of cells containing more than the number of cases expected in the absence of clustering. A large test statistic indicates cluster avoidance such that some of the time intervals have slightly more than the expected number of cases. The test statistic is small when cases cluster in a few time intervals.

A: The test statistic is the number of time intervals with at least $\left\lceil \frac{n}{t} - 0.5 \right\rceil$ cases.

Where:

- *t*: Number of time intervals
- *n*: Total number of cases observed over *t*
- $\frac{n}{t}$: Number of cases expected in an interval in the absence of time clustering
- $\lceil x \rceil$: The least integer greater than *x*. For example, $\lceil 1.3 \rceil = 2$.

According to the null hypothesis, the *n* cases are distributed at random across the *t* time intervals. Under this null hypothesis the expectation and variance are:

$$d = \frac{n}{t} - \left\lceil \frac{n}{t} - 0.5 \right\rceil$$

$$E(A) = t(0.6 + 0.3d)$$

$$Var(A) = 0.155E(A)$$

Significance

A z-score is calculated as:

$$z = \frac{A - E(A)}{\sqrt{Var(A)}},\ z \sim N(0,\ 1)$$

The approximate distribution of z is normal with a mean of 0 and unit variance. P-values are evaluated by comparing z to the percentiles of the normal distribution.

When analyzing several time series simultaneously an overall P-value is obtained as a Chi-squared statistic with one degree of freedom:

$$\chi^2 = \frac{\left(\left| \sum_{i=1}^{s} A_i - \sum_{i=1}^{s} E(A_i) \right| - 0.5 \right)^2}{\sum_{i=1}^{s} Var(A_i)}$$

Dat's Method: How to

Choose Dat's method from the Analysis menu. (Temporal | Single or Several time series submenus).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data file (text or DBF).

   If the text file does not include a label, ClusterSeer will use the row number as the label, which assumes that the sequence of case counts in the file increases with the row number. The DBF file must include a label.

   Your file should contain group-level data with the following columns in the following order:

For single time series:

Time Sequence Label (required for DBF files only)	Case Count

For several time series:

Region Label (required for DBF files only)	Case Count Time 1	Case Count Time 2	Case Count Time 3	...

2. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

3. In the Provide data dialog, you may use the Select File button to change your file choices.
4. Enter the number of time series in your dataset. For a single time series, enter 1. For several time series, enter the total number of locations where you collected data for consecutive intervals of equal length.

Run the analysis

5. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.
6. Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Dat's Method: Results

Plot

To view the plot, choose Plot from the View menu.

ClusterSeer will plot the observed statistic *A* on its expectation *E(A)* for a single and/or several time series. This plot describes where observations would be plotted under the null hypothesis of cases occurring at random over the *t* time periods. The time series are the red points on the graph. The blue line is the identity function (observed=expectation) describing where observations would be plotted under the null hypothesis. Under cluster avoidance $A > E(A)$, and the time series plot above the 45 degree line. When time clustering exists, $A < E(A)$, and time series plot below the line.

Session log

Once ClusterSeer has performed a Dat's analysis, it will place summary information and results into the session log.

Data and analysis input

- Data sets used
- Number of time series
- Number of cells per series
- The test statistic *A*, is the number of cells containing more than the number of cases expected in the absence of clustering.

Results

ClusterSeer reports the results of the analysis for the single and several time series for each row in the time series:

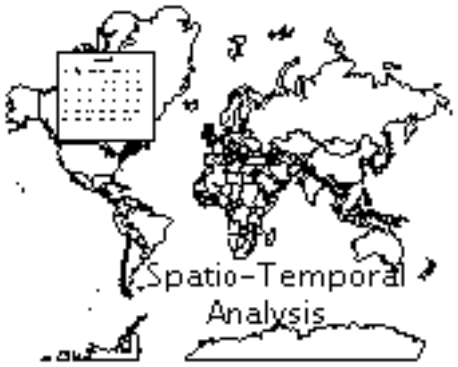
- The expected value of *A*: *(E(A))*
- Its variance: *(Var (A))*
- Z-score
- Upper-tail P-value: The upper-tail P-value is the probability under the null hypothesis of observing an *A* as large or larger than the one given in *A*. It is based on a normal distribution of the data.

For several time series, ClusterSeer also provides:

- A Chi-squared statistic
- An overall P-value
- A list of time series it was unable to analyze.

CHAPTER 19

Direction Method



The Direction method (Jacquez and Oden 1994) tests for space-time interaction of retrospective, individual level data, and calculates the average direction of advance of a spread of cases. The method is sensitive to a systematic, directional spread of cases, such as occurs when an epidemic sweeps through an area. It also arises from geographically localized exposures, with individuals near the source receiving higher doses and showing symptoms before those farther from the source.

You cannot infer a directional process from the Direction method; you can, however, determine whether the observed spread of cases tends to be in one direction.

Direction Method: Statistic

H_o	No association exists between the times at which cases occur and the directions of the vectors formed by connecting the spatial locations of the case.
H_a	The direction from one case to the next is similar for cases that occur at about the same time.

A chain of infection is constructed by first sequencing the cases by time of occurrence. The earliest case would be first, followed by the second case and so on. A line is then drawn to connect the location of the first case to the location of the second case, and this is repeated until all cases are connected. The chain of infection has at least two ends (the first and last cases), and branches when cases occur at exactly the same time.

The test statistic is a vector whose direction is the average direction of the lines composing the chain of infection, and whose magnitude is the angular variance of these marks. When the marks all point in the same direction, the angular variance is small, and when they point in many directions the angular variance is large.

The test statistic is a vector V pointing in the average direction of advance of the chain of infection:

$$v = \frac{1}{m} T \otimes \left(\frac{c}{s} \right)$$

Where:

- c: Cosine matrix whose elements are c_{ij};
- s: Sine matrix whose elements are s_{ij};
- T: Time connection matrix describing the proximity, in time, of the cases to one another.

- $c_{ij} = \cos(\Theta_{ij}) = \left(\frac{\Delta x_{ij}}{\sqrt{\Delta x_{ij}^2 + \Delta y_{ij}^2}} \right)$
- $s_{ij} = \sin(\Theta_{ij}) = \frac{\Delta y_{ij}}{\sqrt{\Delta x_{ij}^2 + \Delta y_{ij}^2}}$
- (x_i, y_i): Geographic coordinates of case *i*.
- Δx_{ij}: Distance on x axis between cases *i* and *j*, $\Delta x_{ij} = (x_i - x_j)$
- Δy_{ij}: Distance on y axis between cases *i* and *j*, $\Delta y_{ij} = (y_i - y_j)$
- Θ_{ij}: Angle between a horizontal line and the vector connecting areas *i* and *j*

The vector *v* points in the average direction of advance of the spread of cases, and its magnitude (termed the angular concentration) represents the variance in the angles between connected cases. When the magnitude is small the variance in the angles is large, and when the magnitude is large the variance in the angles is small. Angles are taken as counter clockwise degrees from horizontal, with East corresponding to 0 and North to 90. Concentration is in the range of 0 to 1, with 1 indicating an angular variance of 0. A consistent direction of spread of cases will result in an angular concentration near 1. A random spread of cases will result in an angular concentration near 0.

The elements of the time connection matrix are determined by the researcher to reflect the suspected temporal scale of the pattern.

ClusterSeer calculates significance values to help you evaluate the significance of the test statistic.

Direction Method: How to

Choose Direction method from the Analysis menu. (Spatio-temporal submenu).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data file (text, DBF, or point shapefile).

   If you submit a text file with a header, a DBF, or a shapefile, ClusterSeer will prompt you to identify which columns in your file contain the required data. In this case, the columns can be in any order.

   If you wish to submit a text file without a header, it should contain individual-level data with the following columns in the following order.

Space-Time Case File:

case level (optional)	case event x-coordinate (not for shapefile)	case event y-coordinate (not for shapefile)	case event time-point (user defined integer)

2. You will need to specify the coordinate system of the data. If the data are in geographic coordinates, you will also need to choose a distance measurement.
3. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

4. In the Provide data dialog, you may use the Select File button to change your file choices.
5. Enter the time connection matrix specification. Suppose N cases occur at times (t_1-t_N). The elements of the time connection matrix are determined by the researcher to reflect the suspected temporal scale of the pattern. See "Alternative directed time measures" on page 200.

Run the analysis

6. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.
7. Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Alternative directed time measures

Use *relative* when you wish to include vectors connecting each case to all of the cases that follow it. This is appropriate when you hypothesize a directional process operating on a longer time span. You are connecting cases that may be several links removed from the chain of infection.

Relative: $t_{ij} = 1$ if $t_j > t_i$; $t_{ij} = 0$ if $t_j = t_i$; $t_{ij} = -1$ if $t_j < t_i$

Use *adjacent* when you wish to connect each case only to its temporal nearest neighbors. These are the cases that occur just before and just after the case. This is appropriate when you hypothesize directional effects of short duration.

Adjacent: $t_{ij} = 1$ if t_j is just after t_i; $t_{ij} = -1$ if t_j is just before t_i; $t_{ij} = 0$ otherwise

Use *following* when you wish to connect each case only to the case (or cases) that immediately follow it. This is appropriate when you wish to trace the average direction of the chain of infection.

Following: $t_{ij} = 1$ if t_j is just after t_i; $t_{ij} = 0$ otherwise

Direction Method: Results

Monte Carlo distribution

You can view the Monte Carlo distribution by choosing MC Distribution from the View menu. The histogram shows the reference distribution generated by randomizing the dataset and recalculating the observed value. The relative position of the observed value of *r* is illustrated with a slim, vertical black line.

Map

You can view the map by choosing Map from the View menu. The map has one layer that shows case-control points.

This map shows a chain of infection. The cases are the black dots, the green lines connect cases sequentially by time of occurrence, and the red lines show the spatial direction of the spread of infection.

If you query one of these points, you can view its label, spatial coordinates, and its time of occurrence.

Session log

After ClusterSeer performs a Direction analysis, it will place summary information and results into the session log.

Parameters and summary statistics

- The file used
- The total number of cases analyzed
- Time measure used (Relative, Adjacent, or Following)

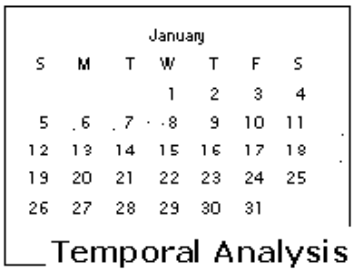
Direction Results

- The vector *r*, the average angle which points in the average direction of advance of the spread of cases. An average angle will always be calculated, whether or not there is a systematic direction to the spread of cases, and is meaningful only when the concentration is statistically significant.

- The angular concentration is the magnitude of *v* and represents the variance in the angles between connected cases.
- The P-value for the angular concentration

CHAPTER 20

Ederer-Myers Mantel Method



The Ederer-Myers-Mantel method (Ederer, Myers, and Mantel 1964) tests for temporal clustering in several time series simultaneously (group level data). Case counts in consecutive time periods for several areas are required. The number of time intervals in the series must be between 2 and 5. The test is insensitive to different population sizes over the areas. Therefore, the method is biased by changes in population size through time.

Examples

Fosgate et al. (2002) studied temporal clustering in human brucellosis in California using the Ederer-Myers-Mantel method. They found significant clustering in Hispanic, non-Hispanic and total cases in different counties in California and in several time periods in 1973-1992. Ward and Carpenter (2000) used this method to evaluate temporal clustering in blowfly strike infestation in Australian sheep flocks.

Ederer-Myers-Mantel Method: Statistic

H_0	Cases occur at random in each time series.
H_a	Cases do not occur randomly through time; they either cluster or occur uniformly.

The test statistic is m_1, the maximum number of cases observed in any of a sequence of time intervals. When cases are clustered m_1 will be large; it will be small when cases occur uniformly through time.

Notation:

- t: number of time intervals
- T: number of time intervals in the time series
- r_i: number of cases in time series i
- $f(r)$: frequency, over all time series, of a given number of total cases
- m_{1i}: the largest number of cases in any time interval of time series i

ClusterSeer constructs a Chi-squared statistic table to test for time clustering in several areas simultaneously. It calculates a Chi-squared test statistic and P-value for each of the following: exact permutation, table values and simulated values.

$$\chi_1^2 = \frac{\left(\left| \sum_{i=1}^{T} m_{1i} - E\left(\sum_{i=1}^{T} m_{1i} \right) \right| - 0.5 \right)^2}{\sum_{i=1}^{T} Var(m_{1i})}$$

The summations are over the number of time intervals in the time series, $\sum_{i=1}^{T} m_{1i}$ is the sum of the maximum number of cases over all time series, $E\left(\sum_{i=1}^{T} m_{1i}\right)$ and $\sum_{i=1}^{T} Var(m_{1i})$ are sums of the expectation and variance of m_{1i} under the null hypothesis.

ClusterSeer calculates several values: estimate and variance values for the exact permutation test, table values, simulated values and simulation P-values for each region in the time series.

There are two ways the distribution of cases can differ from their null distribution: They may cluster (m_1 larger than expected) or they may be uniform (m_1 smaller than expected). Either of these alternatives will inflate the Chi-squared statistic. Use the plot of the expected m_1 on the observed m_1 to determine the nature of the departure (if any) from the null distribution.

Note

- You can request ClusterSeer to report exact values. This calculation may take some time since the process grows exponentially with the number of time series and the number of counts in each time series.
- Whenever possible, use the exact values. When exact values are not available and you must choose between table and simulated values, use simulated, which are generally the most accurate. The simulated and table values should resemble each other.
- Due to the fact that they are approximated, some of the table estimate and variance values are interpolated. ClusterSeer places an asterisk next to the interpolated values in the session log.

Ederer-Myers-Mantel Method: How to

Choose Ederer-Myers-Mantel method from the Analysis menu. (Temporal|Several time series submenus).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data file (text file or DBF).

   If the data file does not include a label, ClusterSeer will use the row number as the label, which assumes that the sequence of case counts in the file increases with the row.

   Your text file or DBF should contain group-level data with the following columns in the following order:

For several time series:

Region Label (required for DBF files only)	Case Count Time 1	Case Count Time 2	Case Count Time 3	...

Choose settings

2. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.
3. Enter the number of time series in your dataset. For a single time series, enter 1. For several time series, enter the total number of locations where you collected data for consecutive intervals of equal length.
4. You may want to check the Calculate True Values Box. If you do, ClusterSeer calculates the exact, rather than interpolated values for the estimate and variance of m_1, and the Chi-squared statistic and P-value. As the number of cases and time cells increases, so does the time it takes to calculate these values. You may press "abort" to stop the calculation process.

Run the analysis

5. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.
6. Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Ederer-Myers-Mantel Method: Results

Plot

You can view the plot by choosing Plot from the View menu. The blue line is the identity function (m_1- $E(m_1)$) describing where observations would be plotted under the null hypothesis. Time clustering causes m_1 to be larger than its expectation, and clustered time series will plot above the dashed line. An equal number of cases in each time interval causes m_1 to be smaller than its expectation, and uniform time series plot below the dashed line.

Session log

After ClusterSeer performs an Ederer-Myers-Mantel analysis, it will place summary information and results into the session log. ClusterSeer reports the file used, the total number of time series, the number of cells per series, and the number of cases in a time series.

Ederer-Myers-Mantel results

- m_1, the test statistic: the maximum number of cases observed in any of a sequence of time intervals
- $E(m_1)$, the expected value of m_1 for the Exact, Table, and Simulated values
- $Var(m_1)$, the variance of m_1 for the Exact, Table, and Simulated values
- P-values for each region in the time series. Exact if available, simulated if not
- The interpolated values of $E(m_1)$ and $Var(m_1)$ (identified with an asterisk)
- The number of Monte Carlo simulations
- A list of series ClusterSeer was unable to analyze

Chi-squared statistic results

True Values

- The value of the Chi-squared statistic of the True values (if you selected Calculate True values in the Provide data to run the Ederer-Myers-Mantel test dialog)
- The P-value for the Chi-squared statistic of the True values

Table Values

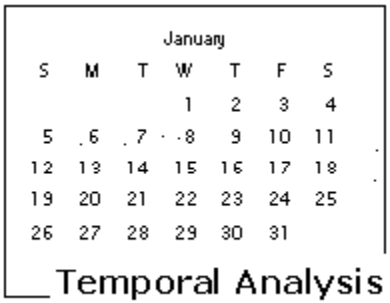
- The value of the Chi-squared statistic of the Table values
- The P-value for the Chi-squared statistic of the Table values

Simulated Values

- The value of the Chi-squared statistic of the simulated values
- The P-value for the Chi-squared statistic of the simulated values

CHAPTER 21

Empty Cells Method



The Empty cells method (Grimson 1993) can detect temporal clustering in single and several time series with group-level data. The empty cells test statistic, *E*, is a count of the number of time periods with no cases. This statistic can detect temporal clustering where one or more time periods have several cases while other time periods have none. When cases cluster the test statistic will be large.

Use this method to detect clusters of rare events, when some of the time periods can be reasonably expected to have zero cases.

Empty Cells Method: Statistic

H_o	Cases occur randomly through time.
H_a	Cases cluster in one or more time periods.

Grimson (1993) gives equations for the expectation and variance under this null hypothesis as:

$$E(E) = t\left(\frac{t-1}{t}\right)^N$$

$$E((E)_2) = (t)_2 t^{-N}(t-2)^N$$

$$Var(E) = E(E)(1-E(E)) + E((E)_2)$$

Where:

- N: Number of cases in a time series
- t: Number of time cells
- E: Test statistic, number of empty cells.
- The notation $(a)_k$ indicates a falling factorial such that $(a)_k=a(a-1)...(a-k+1)$.

Significance

Under this null hypothesis we wish to determine the probability, P, of obtaining a number of empty cells greater than or equal to E. The significance of E is evaluated using the exact P-value:

$$P(E \geq E^*) = (-1)^{E^*} \sum_{k \geq E^*}^{t-1} (-1)^k \binom{k-1}{E^*-1} \binom{t}{k} \left(\frac{t-k}{t}\right)^N$$

The notation $\binom{a}{b}$ indicates a binomial coefficient. We want to test for clusters, and $P(E \geq E^*)$ is evaluated as a one-tailed test.

When several time series are tested simultaneously P-values are combined using the Bonferroni approach. When at least 20% of the areas have an expected number of empty cells of 5 or more the results can be combined as a continuity-corrected Chi-squared statistic with one degree of freedom.

$$x^2 = \frac{\left(\left| \sum_{i=1}^{t} E_i - \sum_{i-1}^{t} E(E_i) \right| - 0.5 \right)^2}{\sum_{i=1}^{t} Var(E_i)}$$

Here the *i* subscript indicates a statistic for the *i*th time series, the summations are over the cells/time periods within each series, *E* is the sum of the number of empty cells, and *E(E)* and *Var(E)* are the mean and variance of *E* under the assumption of a random allocation of disease cases among the cells.

Empty Cells Method: How to

Choose Empty Cells method from the Analysis menu. (Temporal | Single or Several time series submenus).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data file (text file or DBF).

   DBFs must have a column of labels. If you submit a text file without a label column, ClusterSeer will use the row number as the label, which assumes that the sequence of case counts in the file increases with the row number.

   The text or DBF file should contain group-level data with the following columns in the following order:

For single time series:

Time Sequence Label (required for DBF files only)	Case Count

For several time series:

Region Label (required for DBF files only)	Case Count Time 1	Case Count Time 2	Case Count Time 3	...

2. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

3. In the Provide data dialog, you may use the Select File button to change your file choices.
4. Enter the number of time series in your dataset. For a single time series, enter 1. For several time series, enter the total number of locations where you collected data for consecutive intervals of equal length.

Run the analysis

5. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.
6. Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Empty Cells Method: Results

Plot

Choose Plot from the View menu to view the plot. ClusterSeer plots the observed statistic (E) on its expectation ($E(E)$) for single and several time series. The time series is the single red point on the graph. The blue line is the identity function (observed = expectation) describing where observations would be plotted under the null hypothesis. Significant clustering will cause E to be larger than its expectation, and observations will plot above the blue line.

Session log

After ClusterSeer performs an Empty Cells analysis, it will place summary information and results into the session log.

ClusterSeer also provides a table of the test statistic, the count of empty cells E, and a list of time series ClusterSeer was unable to analyze.

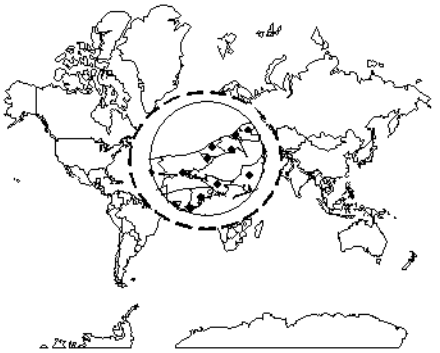
Significance Values

- The expectation of the number of empty cells $E(E)$
- The variance of the number of empty cells $Var(E)$
- The upper-tail P-value to assess the significance of the observed value of E
- Overall Bonferroni P-value or Chi-squared statistic with a P-value (for multiple time series only)

When several time series are tested simultaneously P-values are combined using the Bonferroni approach. When at least 20% of the areas have an expected number of empty cells of 5 or more, the results can be combined as a continuity-corrected Chi-squared statistic with one degree of freedom.

CHAPTER 22

Getis-Ord Local G



The local *G* statistics are used to test for spatial clustering in group-level data (Getis and Ord 1992; Ord and Getis 1995). These statistics make it possible to assess the spatial association of a variable within a particular distance of each observation. Local *G* statistics may detect local clusters that exist despite negative tests for global spatial autocorrelation.

Examples

Jeffery et al. (2002) used the Local *G* statistic to examine the spatial pattern of mosquito vectors of the Ross Rover virus and the Barmah Forest virus on Russell Island, Queensland, Australia. They found significant clustering on the southern end of Russell Island. Ratcliffe and McCullagh (2001) located hotspots in crime in Nottinghamshire using the Local *G* statistic. They then compared statistical clusters with the perceived high crime areas. Ceccato and Persson (2002) used the Local *G* to study spatial clustering of employment in Sweden. They found clusters of employment in areas with private businesses, and clusters of low employment in areas where the government is the main employer.

Getis-Ord Local G Method: Statistic

H_o	There is no clustering of high or low values within the specified distance of location *i*, and the test statistic is close to zero.
H_a	There is clustering of high or low values within the specified distance of point *i*. A significant positive value implies a clustering of high values, and a significant negative value indicates a clustering of low values.

The statistic measures the degree of association that results from the concentration of weighted points or region centroids and all other weighted points within distance *d* from the point of study (Getis and Ord 1992).

The basic statistic is defined as:

$$G_i(d) = \frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j}$$

In this equation, the x_j are the weighted values of the points in the study area. w_{ij} is a binary, symmetric weights matrix with ones for all points *j* within distance *d* of point *i* and zeros otherwise.

There are two variants of the local *G* statistic. The G_i statistic excludes the value at *i* from the summation and is used for spread or diffusion studies, while the G_i^* includes the value at *i* in the summation and is most often used for studies of clustering.

Significance

In (Ord and Getis 1995) the authors reformulated the statistic so that the results are given in standard normal variants. The statistic is normally distributed, and can be used for normal as well as skewed frequency distributions of the underlying variable. However, when the number of neighbors is small the statistic is less reliable.

The significance of each local *G* value may also be evaluated using a Monte Carlo randomization procedure.

Getis-Ord Local G Method: How to

Choose Getis-Ord Local *G* method from the Analysis menu. (Spatial | Local submenus).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data file (text, DBF, or point shapefile).

   If you submit a text file with a header, a shapefile, or DBF, ClusterSeer will prompt you to identify which columns in your file contain the required data. In this case, the columns can be in any order.

   If it is a text file without a header, it should contain individual-level data with the following columns in the following order.

centroid label	centroid x-coordinate (not for shapefile)	centroid y-coordinate (not for shapefile)	disease frequency

2. You will need to specify the coordinate system of the data. If the data are in geographic coordinates, you will also need to choose a distance measurement.
3. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

4. In the Provide data dialog, you may use the Select File button to change your file choices.
5. Select the distance of study. When a new dataset is selected, the initial distance shown in the dialog is the largest nearest neighbor distance in the study area. This ensures that each point will have at least one neighbor in the analysis. Any distance may be chosen; however, the selected distance should be based on a hypothesis about the nature of the clustering or diffusion forces underlying the data.
6. Select either a G_i or G_i^* test.
7. Choose the number of Monte Carlo runs, the number of simulations used to determine statistical significance of the test statistic. The default value is 999.

8. Enter the significance level you wish to use for the test. The significance level is the alpha level, the cutoff for statistical significance. The default value is 0.05.

Run the analysis

9. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.
10. Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Getis-Ord Local G Method: Results

Distribution

You can view a histogram that shows the reference distribution from the Monte Carlo simulations. ClusterSeer has a Monte Carlo distribution for each region in your dataset.

Choose MC Distribution from the View menu. Next, ClusterSeer will prompt you to choose a region from the list of regions in your dataset.

The distribution of test statistics from the simulations will appear as gray bars, and the observed test statistic will be drawn as a slim black line.

Map

Choose Map from the View menu to view the map. The map has three layers: clusters of high values, region centroids (observations), and a map showing the selected distance for all points

If you query a region centroid, you will be able to view its label, x, and y coordinates, disease frequency, local test statistic, normal P-value, and Monte Carlo P-value at the point.

Layer	Q?
clusters of high values	If you query a cluster layer, you can view the center area label, center x, y coordinates, local test statistic, normal P-value, Monte Carlo P-value, and the local disease frequency.

Layer	Q?
region centroid	When you query the region centroid, you can view its label, x, and y coordinates, disease frequency, local test statistic, normal P-value, and Monte Carlo P-value at the point.
distance layer	The same attributes as the cluster layer can also be found for each point by querying a circle in the selected distance layer.

Session log

After ClusterSeer performs a Getis-Ord analysis, it will place summary information and results into the session log.

Summary Information and Parameters:

Where applicable, ClusterSeer reports conversion information from geographic to planar coordinate systems. It also reports the total number of regions and the average disease frequency.

Local G_i or G_i Test Summary information:*

- File name
- Coordinate system conversion information
- Total number of regions
- Specified alpha level
- Number of Monte Carlo simulations
- Distance of analysis
- Locations with significant clustering and the corresponding test statistic
- P-value
- Monte Carlo P-value

CHAPTER 23

Grimson's Method

Grimson's method (Grimson 1989, Grimson and Rose 1991) is a versatile test used to detect space, time or space-time clustering in time series and/or point data. It can be used with individual-level or group-level data. Grimson's method is sensitive to a high number of adjacent high risk events. The data may consist of items labeled cases for individual point data, high risk areas for group-level data, or high-risk time periods (e.g. when an exposure or diagnosis occurred) for time series data.

Grimson's Method: Statistic

H_o	The objects have been labeled at random, there is no clustering.
H_a	High risk items tend to be adjacent, there is clustering.

Test Statistic

The test statistic, *A*, is the count of the number of pairs of labeled objects that are adjacent to one another in a time series or in space, or both. The objects can be locations of cases and controls, high risk areas or time periods. Objects may border each other in space (sharing a common boundary), or time (adjacent in a time series). The expected value of *A* when there is no clustering of events is:

$$E(A) = \frac{yn(n-1)}{2(x-1)}$$

Here *x* is the total number of items (both labeled and not labeled), *n* is the number of labeled items, and *y* is the average number of borders per item (neighbors in time and/or space). When high-risk items cluster there will be an excess of adjacencies and the test statistic will be large.

The variance of *A* is:

$$Var(A) = E(A)\left(1 + \frac{2(y-1)+(n-2)}{x-2} + \frac{(xy-4y+2)(n-2)(n-3)}{2(x-2)(x-3)} - E(A)\right) +$$

$$Var(y)\left(\frac{(n)_3}{(x-1)_2} - \frac{(n)_4}{(x-1)_3}\right)$$

The variance of A has two components, the regularity component (RC) and the variability component (VC). RC is the first part of the expression (RC = E(A)[....]) and VC is the second part (VC = Var(y)....).

Significance

ClusterSeer evaluates the significance of *A* using the Poisson or the normal distribution. The first assumes *A* is sampled from a Poisson distribution with a mean given by $\boldsymbol{E}(A)$. The second assumes the z-score is sampled from a normal distribution with mean of 0

and variance 1.0. Both approaches yield a one-tailed test describing the probability, under the null hypothesis, of obtaining a test statistic as large or larger than the one already observed.

Which distribution is right for my data?

Whether to use the Poisson or the normal approach depends on the proportion of the variance, *Var(A)*, contributed by the variability component, *VC*. This is `*VC* / *Var(A)*'. Grimson (1991) offers the following guidance:

- Use the Poisson approach when *VC* / *Var(A)* is small
- Use the normal approach when *VC* / *Var(A)* is large.

Our rule of thumb is to use the Poisson approach when *VC* / *Var(A)*< 0.20, otherwise use the normal approach. The mean and variance of the Poisson distribution are equal, thus *E(A)* and *Var(A)* should be approximately equal in order to use the Poisson approach.

Grimson's Method: How to

Choose Grimson's method from the Analysis menu. As it is a flexible test, it can be found under a series of submenus: (Spatial|Global) or (Temporal|Single or Several) or (Spatio-temporal).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

ClusterSeer can either test parameters you enter in the dialog or it can calculate the necessary parameters from data files you submit. To use the files, you will need to submit two files: a data file (a text file), and a file of contiguities or adjacencies (text or gal file). See pp. 42-44 of the first ClusterSeer user guide for information on text and contiguity file guidelines.

Enter parameters directly

You can calculate the parameters needed to run Grimson's method outside of ClusterSeer and then enter them into the dialog box that appears. Because of the complexity of the variance calculation, submitting the files to allow ClusterSeer to calculate the parameters from a file may be the easier approach.

Then, enter the following parameters directly into the dialog:

- Total number of objects
- Number of labeled objects
- Average number of borders
- Variance of the number of borders
- Number of adjacent labeled pairs

OR

Enter your parameters automatically using file information

1. Use the Select file button to choose your files.

Grimson: Disease data file

2. For Grimson's method, you can submit two files for ClusterSeer to calculate the parameters used in Grimson's method. You will need to submit a disease data file and a contiguity file. For the disease file, ClusterSeer is flexible. It can accept two types of disease data: risk labels and case counts.

Risk label

3. For the risk label files, you will submit a text file that holds two columns of data: the label (for the location or time period) and the risk label (0 for low risk, 1 for high risk).

Case count

4. ClusterSeer will calculate the risk labels for analysis from case count data. Use the dialog to define what you consider a high risk case count. You can set it to any cell with a case count equal to (=), not equal to (not =), less than (<), greater than (>), less than or equal to (<=), or greater than or equal to (>=) a specified value. Objects are labelled high risk if they have a case count of 2.

Grimson's Method: Results

Plot

To view the plot, choose Plot from the View menu.

The plot shows the significance of *A* against the value of *A*. The Poisson significance is shown by the red points. The blue points are significance under the normal approach. The vertical green line is the observed number of adjacent high-risk quarters. The intersection between the Poisson and normal curves and the vertical line are the P-values under the Poisson and normal assumptions. The P-values are different for the Poisson and normal distributions. See "Which distribution is right for my data?" on page 227.

Session log

Once ClusterSeer has performed a Grimson's analysis, it writes information on the analysis and results into the session log.

Data and analysis input

- Data sets used
- Number of objects
- Number of labelled cells
- Average number of borders per cell
- Sample variance for number of borders
- The test statistic *A*, the number of adjacent labelled cells.

Results

ClusterSeer will then report the results of the analysis. These include the variance components *RC* and *VC* calculated from the data, the expected value of *A (EA)*, its variance, and the z-score.

ClusterSeer reports two upper-tail P-values, one obtained from the normal approach and the other from the Poisson approach. Use the Poisson approach when *VC/Var(A)* is small, and the normal approach when *VC/Var(A)* is large (Grimson 1991). One rule of thumb is to use the Poisson approach when *VC/Var(A)*<0.20; otherwise, use the normal approach. See "Which distribution is right for my data?" on page 227.

CHAPTER 24

Jacquez's k- NN

Jacquez's *k*-Nearest Neighbor test is a test for space-time interaction for individual-level data. The test statistic, J_k, is the count of the number of case pairs that are nearest neighbors in both space and time. When space-time interaction exists J_k will be large, since nearest neighbors in space will also tend to be nearest neighbors in time.

Examples

Norstrom et al. (2000) used the *k-NN* method to evaluate outbreaks of acute respiratory diseases in Norwegian cattle herds. They found significant space-time clustering of cases, providing evidence of a single infection source for the outbreak. Van Buuren et al. (1998) found no significant space-time clustering of multiple sclerosis cases in The Netherlands using the *k-NN* method.

Jacquez's k-NN Method: Statistic

H_o	Whether cases are nearest neighbors in space is independent of whether they are nearest neighbors in time.
H_a	Nearest neighbors in space tend to be nearest neighbors in time.

Test statistic

The test statistic, J_k, is the count of the number of case pairs that are k nearest neighbors in both space and time. When space-time interaction exists J_k will be large, since nearest neighbors in space will also tend to be nearest neighbors in time.

$$J_k = \sum_{i=1}^{N} \sum_{i=1}^{N} s_{ijk}\, t_{ijk}$$

Where:

- k is the number of nearest neighbors to consider in the analysis (if k=1, consider the first nearest neighbor; if k=2 consider the first and second nearest neighbors).
- s_{ijk} is the spatial nearest neighbor (*NN*) measure, s_{ijk} =1 if case j is a k-*NN* of case i in space, and 0 otherwise.
- t_{ijk} is the time *NN* measure, t_{ijk} =1 if case j is a k-*NN* of case i in time, and 0 otherwise.

The J_k are not independent because case pairs counted as nearest neighbors when k is small are included when higher numbers of neighbors are considered. For example, J_2 is the count of the number of pairs that are first and second nearest neighbors in both space and time. DJ_k is the number of space-time nearest neighbors added by increasing k by 1. DJ_k measures space-time interaction above and beyond that observed for the k-1 nearest neighbors. J_k, on the other hand, is a cumulative measure of space-time interaction where all nearest neighbor relationships from 1 up to k are included.

$$DJ_k = J_k - J_{k-1}$$

Significance

ClusterSeer evaluates the significance of J_k using an approximate randomization of the Mantel product. Let S_k denote the matrix of spatial nearest neighbor measures (the s_{ijk}), and let T_k denote the matrix of time nearest neighbor measures. P-values are calculated by comparing the observed J_k to the distribution of J_k obtained under Monte Carlo randomization. The elements of T_k are shuffled by permuting its rows and columns and J_k is then calculated. This procedure is repeated a fixed number of times, resulting in a distribution of J_k under the null hypothesis of no association between the space and time nearest neighbor relationships.

Jacquez's k-NN Method: How to

Choose Jacquez's *k*-Nearest Neighbor Method from the Analysis menu (Spatio-temporal submenu).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data file (text file, DBF, or point shapefile). ClusterSeer expects case event timepoints as user-defined integers. See "Temporal data formats" on page 174.

   If you submit a text file with a header, a shapefile, or DBF, ClusterSeer will prompt you to identify which columns in your file contain the required data. In this case, the columns can be in any order.

   If you wish to submit a text file without a header, it should contain individual-level data with the following columns in the following order

Space-Time Case File:

case level (optional unless you are submitting a DBF)	case event x-coordinate (not for shapefile)	case event y-coordinate (not for shapefile)	case event timepoint (user-defined integer)

2. You will need to specify the coordinate system of the data. If the data are in geographic coordinates, you will also need to choose a distance measurement.
3. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

4. In the Provide data dialog, you may use the Select File button to change your file choices.
5. Enter the number of case pairs that are *k*-Nearest Neighbors in both space and time.
6. Enter the number of Monte Carlo Randomization runs.

Run the analysis

7. After you hit OK, ClusterSeer will establish nearest neighbor relationships. Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Jacquez's k-NN Method: Results

Monte Carlo distribution

You can view the Monte Carlo distribution by choosing MC Distribution from the View menu.

This histogram shows the reference distribution generated by randomizing the dataset and recalculating the observed value. The relative position of the observed value of J_k is illustrated by a slim, vertical black line.

Map

Choose Map from the View menu. ClusterSeer will display a map of the cases' spatial distribution.

Q? If you query one of these points, you will be able to view its label, spatial coordinates, and its time of occurrence.

Session log

After ClusterSeer performs a Jacquez's *k*-Nearest Neighbor analysis, it will place summary information and results into the session log.

- The file used
- The number of cases analyzed

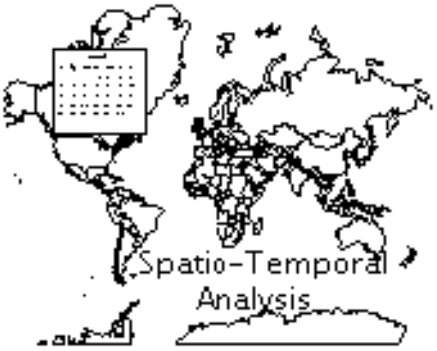
Jacquez's *k-NN* results

ClusterSeer will provide a table of results, including the test statistics J_k and DJ_k at each *k* and P-values for each test statistic obtained by Monte Carlo randomizations.

ClusterSeer also provides the Statistical Distance Test Statistic for the combined Monte Carlo simulation runs at each *k*, its P-value, and Bonferroni and Simes combined P-values for *J* and *DJ* over all *k* analyzed.

CHAPTER 25

Knox's Method



Knox's method (Knox 1963, 1964) quantifies space-time interaction for individual-level data. The test statistic, *X*, is a count of those pairs of cases that are separated by less than the critical space and time distances. Pairs of cases will be near to one another when interaction is present, and the test statistic will be large.

Example

Gilman et al. (1999) assessed patterns in acute lymphoblastic leukemia in the UK using the Knox method. They found space time clustering only in children 0-14 years, particularly those diagnosed in 1984-88. This pattern could be explained by an infectious etiology or due to exposures to an environmental hazard. Machado-Coelho et al. (1999) used the Knox test to study American cutaneous leishmaniasis (ACL) in Brazil. They found significant space-time pattern in ACL cases.

Knox's Method: Statistic

H_o	The times of occurrence of the health events are distributed randomly across the case locations. This is another way of saying the time distances between pairs of cases are independent of the spatial distances between pairs of cases.
H_a	Pairs of cases near in space tend to be near in time.

Test statistic

The test statistic, *X*, is the number of pairs of cases that are near to one another in both space and time. Pairs of cases will be near to one another when interaction is present, and the test statistic will be large.

$$x = \sum_{i=1}^{N} \sum_{j=1}^{i-1} s_{ij}\, t_{ij}$$

Where:

- *N* is the number of cases
- s_{ij} is the space adjacency value, 1 if the distance between cases *i* and *j* is less than the critical space distance and 0 otherwise
- t_{ij} is the time adjacency value, 1 if the waiting time between cases *i* and *j* is less than the critical time distance and 0 otherwise.

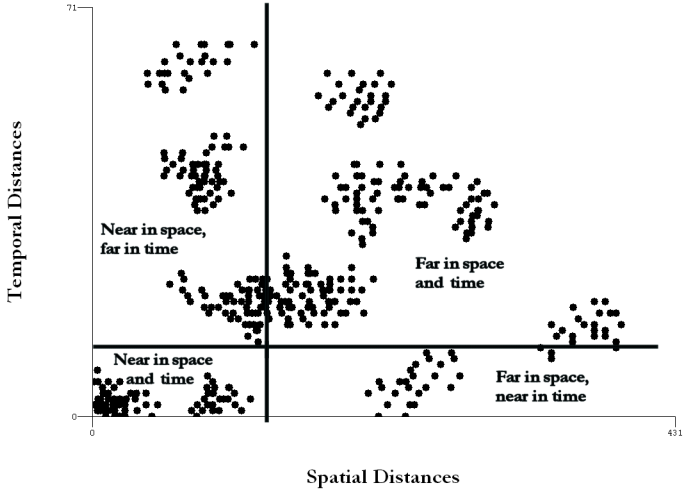
To run Knox's method, you need to specify the critical space and time distances (*Dcrit* and *Tcrit* respectively), so that s_{ij} and t_{ij} can be calculated. Pairs of cases separated by less than the critical space distance are considered to be near in space. Pairs of cases separated by less than the critical time distance are said to be near in time. If you do not know what critical distances to use, ClusterSeer will use the mean calculated from the dataset.

Significance

ClusterSeer calculates the null distribution of *X* in two ways: using a Chi-squared test and using Monte Carlo simulations. The Chi-squared test calculates the probability of the classification of events into near in space and near in time; near in space, far in time; far in space, near in time; and far in space and time under the null hypothesis of no clustering. This provides a significance based on a comparison of the observed and expected values of *X*.

Critical values

Knox's method detects clustering using threshold values, critical time and space distances. Pairs of cases separated by less than the critical space distance are considered to be near in space. Pairs of cases separated by less than the critical time distance are said to be near in time. To run Knox's method, you need to supply these critical values.



The plot above (not from ClusterSeer) shows the time critical value as a horizontal line, the spatial distance critical value as a vertical line. Pairs of events are categorized as near or far in space and time. There are 4 categories or classifications, as shown on the plot.

How do you determine when cases of a disease are "close enough" in time or space?

Knox designed this method to account for latency periods. A latency period is the time between exposure and the manifestation of symptoms. If you suspect a disease with a latency period of 3 days set the time critical distance long enough to allow symptoms to appear, say 4 or 5 days. For infectious diseases, the geographic critical distance reflects

the average distance between 2 individuals, one of whom infected the other. In general, one selects critical distances consistent with the disease hypothesis under investigation. This hypothesis based approach avoids problems of subjectivity which arise when critical values are determined from the data.

However, when knowledge of the underlying disease process is absent, critical values can be quantified based on the distributions of space and time distances. This approach is crude and should only be used when an epidemiologic hypothesis is lacking. In these instances use the mean geographic distance for *Dcrit* and the mean time distance for *Tcrit*. You can systematically vary the critical distances to identify those values that maximize Knox's *X*. This can provide insight into the spatial and temporal scale of the disease process, but precludes any formal evaluation of statistical significance because of multiple tests. Do not choose critical distances larger than the maximum distance in the data, since the number of cases near in both space and time will always be zero.

Mantel's method is another space-time test which does not require selecting critical distances.

Knox's Method: How to

Choose Knox's Method from the Analysis menu (Spatio-temporal submenu).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

This analysis requires a single file of case event data. Files should follow ClusterSeer general data requirements. ClusterSeer expects case event timepoints as user-defined integers. See "Temporal data formats" on page 174.

1. ClusterSeer will prompt you to submit the data file (text file, DBF, or point shapefile).

   If you submit a text file with a header, a shapefile, or DBF, ClusterSeer will prompt you to identify which columns in your file contain the required data. In this case, the columns can be in any order.

   If you wish to submit a text file without a header, it should contain individual-level data with the following columns in the following order

Space-Time Case File:

case level (optional unless you are submitting a DBF)	case event x-coordinate (not for shapefile)	case event y-coordinate (not for shapefile)	case event timepoint (user-defined integer)

2. You will need to specify the coordinate system of the data. If the data are in geographic coordinates, you will also need to choose a distance measurement.
3. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

4. In the Provide data dialog, you may use the Select File button to change your file choices.
5. Specify the distance and temporal cutoffs. Select whether or not you want to use means as cutoffs. See "Critical values" on page 239.
6. Enter the number of Monte Carlo Randomization runs.

Run the analysis

7. After you hit OK, ClusterSeer will establish nearest neighbor relationships. Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Knox's Method: Results

Monte Carlo distribution

You can view the Monte Carlo distribution by choosing MC Distribution from the View menu.

This histogram shows the reference distribution generated by randomizing the dataset and recalculating the observed value. The relative position of the observed value of *X* is illustrated with a slim, vertical black line.

Map

You can view the map by choosing Map from the View menu.

The map of Knox results has seven layers, one layer showing the original data (cases) and six showing the analysis results. When the map is first displayed, not all of the layers are active (checked in red on the left layers pane and shown in the middle pane of the map). You can toggle map layers on and off by checking or clearing the red check using the mouse pointer.

Layer type	Layer	Contents
Point Layer	Close in Space and Time	Those points that are close in space and time to at least one other point
	Close in Time	Those points that are close in time to at least one other point
	Close in Space	Those points that are close in space to at least one other point
	Cases	The point locations of all cases in the dataset
Link Layers	Space-Time Links	The links between pairs of points that are near in space and time
	Time Links	The links between pairs of points that are near in time
	Space Links	The links between pairs of points that are near in space

Q? To query a particular layer, make sure it is highlighted on the layers list (left map panel).

- If you query a point layer (cases, close in space, close in time, or close in space and time), you will see information on the point nearest the query location. The label, x-coordinate, y-coordinate, and the time of its occurrence.
- If you query a links layer (space links, time links, or space-time links), you will see the layer name and the coordinates of the location queried.

Session log

After ClusterSeer performs a Knox analysis, it will place summary information and results into the session log:

- The file used,
- The number of cases analyzed,
- The critical values, *Dcrit* and *Tcrit,*
- A 2x2 matrix categorizing pairs of cases in the dataset as close or far in space and time.

Chi-square results

ClusterSeer reports the expected test statistic and the Chi-squared P-value.

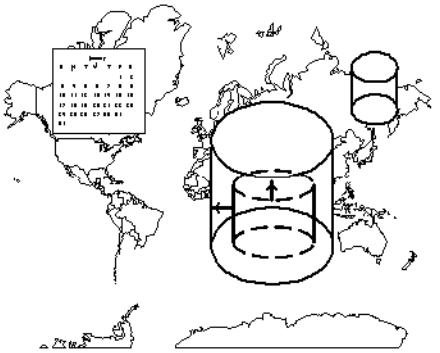
The Chi-squared test calculates the probability of the classification of events into near in space and near in time; near in space, far in time; far in space, near in time; and far in space and time under the null hypothesis of no clustering. This provides a significance based on a comparison of the observed and expected values of *X*.

Monte Carlo results

- The test statistic *X*
- The number of Monte Carlo simulations
- The P-value for the test statistic through comparison with the Monte Carlo distribution.

CHAPTER 26

Kulldorff's Spatial Scan



Kulldorff's Spatial Scan method (Kulldorff and Nagarwalla 1995, Kulldorff 1997) can detect local space clusters in group-level data. The first ClusterSeer manual page 75 describes Kulldorff's Spatio-Temporal scan. ClusterSeer 2 now offers the space-only version. The scan statistic uses a circular window to identify excesses of cases in space. At each spatial location, a circular window increases in size until it reaches an upper size limit.

The scan statistic provides a measure of whether the observed number of cases is unlikely for a window of that size, using reference values from the entire study area. By searching for clusters without specifying their size or location, the method avoids pre-selection bias.

Kulldorff (1997) developed two models, a Poisson model and a Bernoulli model. For a small number of cases, the two models are similar. The Bernoulli model is best for questions about case and control samples, while the Poisson model better answers questions with case and population-at-risk counts. At this point, ClusterSeer implements the Poisson method.

Examples

Doherr et al. (2002) used the spatial scan statistic to analyze spatial clustering in cases of bovine spongiform encephalopathy (BSE) in Switzerland. They found significant spatial clustering in BSE cases, and they excluded differential reporting as a possible cause of the observed clustering. Ward and Carpenter (2000) used the spatial scan statistic to identify patterns in blowfly strike infestation in commercial sheep flocks in Australia.

Kulldorff's Spatial Scan Method: Statistic (Poisson)

H_0	The null spatial model is an inhomogeneous Poisson point process with an intensity, μ, proportional to the population-at-risk.
H_a	In some locations in the space, the number of cases exceeds that predicted under the null model.

Test statistic

For the spatial scan, a circular window is moved systematically through the study area. The scan window starts at each location in the dataset. ClusterSeer's implementation of the spatial scan calculates the locations to include in the window using the centroids of a submitted polygon file. The window expands to include the nearest region centroids. The maximum size of each window will not exceed 50% of the total population-at-risk size for the study period.

The hypotheses are evaluated with a maximum likelihood ratio test that examines whether the null or alternative model better fits the data (notation follows Kulldorff 1999). The scan statistic is the maximum likelihood ratio over all possible window sizes. Its P-value is obtained through multinomial Monte Carlo randomizations. If the null hypothesis is rejected, ClusterSeer reports the spatial or spatio-temporal location and the extent of the cluster that caused the rejection.

Likelihood ratio

The likelihood ratio is

$$\frac{L(Z)}{L_0} = \frac{\left(\frac{n_z}{\mu(Z)}\right)^{n_z}\left(\frac{N-n_z}{N-\mu(Z)}\right)^{N-n_z}}{\left(\frac{N}{\mu(A)}\right)^{N}}$$

if $n_z > \mu(Z)$, $\frac{1}{L_0}$ otherwise

Where n_z is the observed number of cases and $\mu(Z)$ is the expected number of cases in cylinder *Z*. The observed (*N*) and expected $\mu(A)$ number of cases are calculated over the entire study area, across all time periods.

Kulldorff's Spatial Scan Method: How to

You can perform a Kulldorff's Scan in one of three ways, submitting population-at-risk counts directly with case counts, extrapolating population-at-risk counts from census data, or submitting a shapefile of polygons with case and population-at-risk counts.

Only the method new to ClusterSeer version 2 is described here. See the first ClusterSeer user guide (pp. 77-80) for a description of how to submit population at risk counts or census data.

Note: For the polygon shapefile, if there are islands or non-contiguous areas, the analysis will not run.

Submit shapefile

Choose Kulldorff's Scan method from the Analysis menu (Spatial|Local submenus).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

1. ClusterSeer will prompt you to submit a point or polygon shapefile, text or DBF file.

   If you submit a shapefile, DBF, or text file with a header, ClusterSeer will prompt you to identify which columns in your file contain the required data. In this case, the columns can be in any order.

   If you wish to submit a text file without a header, it should contain group-level data with the following columns in the following order:

region label	region centroid x-coordinate (not for shapefile)	region centroid y-coordinate (not for shapefile)	case count	population-at-risk count

2. You will need to specify the coordinate system of the data. If the data are in geographic coordinates, you will also need to choose a distance measurement.
3. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

4. In the Provide data dialog, you may use the Select File button to change your file choices.

5. Choose the number of Monte Carlo runs, the number of simulations used to determine statistical significance of the test statistic. The default value is 999.

Run the analysis

6. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.

Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Kulldorff's Spatial Scan Method: Results

Monte Carlo distribution

This histogram shows the reference distribution generated by randomizing the dataset and recalculating the test statistic. See p. 80 of the first ClusterSeer manual for details on Spatial and Spatio-Temporal Kulldorff's Scan distributions.

Map

To see the map, choose Map from the View menu. The map will display two layers: region centroids, shown as points, and cluster extent, shown as a circular outline for each of the three most likely clusters. The second and third most likely clusters are chosen using two criteria: 1) the value of the test statistic and 2) whether they overlap higher-ranking clusters (the second will not overlap the first, the third will not overlap the second or the first).

Q? To query a particular layer, make sure it is highlighted on the layers list (left map panel).

Layer	Q? For the Spatial Scan	Q? For the Space-Time Scan
region centroids	You can view the region label, case count, and population at risk count.	You can view the region label, x- and y- coordinates, case count, and population at risk count.
cluster extent	You can find its centering region label, x- and y- coordinates, case count and population count.	You can find its centering region label, x- and y- coordinates, start and end periods for the cluster, local test statistic, disease frequency, P-value, and a list of other regions included in the cluster.

Plot

The spatial scan has no plot option. See page 81 of the first ClusterSeer manual for information about the spatio-temporal scan's plot.

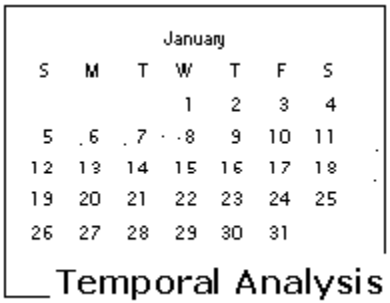
Session log

Once ClusterSeer has performed a Kulldorff's Scan analysis, it writes information on the procedure and results into the session log.

Session Log Topics	For the Spatial Scan	For the Spatio-Temporal Scan
Summary Information & Parameters	• Number of regions, number of cases, population-at-risk size, average disease frequency • Variable analyzed (case count, population count), maximum spatial population radius analyzed, number of Monte Carlo simulations performed	• Number of regions, study period span, number of cases, population-at-risk size, average disease frequency • Maximum population radius, maximum temporal span, number of Monte Carlo simulations
Information on each of the three most likely clusters: The second and third most likely clusters are chosen using two criteria: 1) the value of the test statistic and 2) whether they overlap higher-ranking clusters (the second will not overlap the first, the third will not overlap the second or the first).	• Regions included (starting with the centering region, with remaining regions ordered from nearest to farthest) • Disease frequently (averaged over the spatial span) • Log likelihood ration • Upper tail Monte Carlo P-value	• Regions included (starting with the centering region, with remaining regions ordered from nearest to farthest) • Cluster temporal span • Disease frequency (averaged over the cluster temporal span) • Log likelihood ratio • Upper tail Monte Carlo P-value

CHAPTER 27

Larsen's Method



Larsen's method can detect temporal clustering in single and several time series with group-level data. Larsen's test statistic *K* is sensitive to a unimodal clustering of occupied cells.

The normality assumption used to evaluate significance does not hold when the time series is shorter than 10 intervals. When you use Larsen's method to analyze time series with fewer than 10 intervals, the method will not be very powerful. For short time series consider using smaller time intervals or perhaps collecting data from additional time periods. As alternatives consider the Empty Cells method, or Dat's 0-1 matrix method when cases are numerous.

Larsen's method also requires two or more of the time intervals to have cases. Time series with fewer than 2 occupied intervals are excluded from an analysis. Each time series must have at least 1 unoccupied cell.

You cannot use Larsen's method with rates. Counts are required and the method is biased by changes in population size through time.

Example

Larsen et al. (1973) used the method for simultaneous clustering in several time series to screen for clusters of cases of children with acute leukemia in 18 census tracts in Atlanta, Georgia. Attention was brought to these data

because the leukemia within census tracts appeared to cluster. Upon further scrutiny, single leukemia clusters were identified in 10 of the tracts. The data showed significant unimodal clustering when all of the census tracts were considered simultaneously.

Larsen's Method: Statistic

H_o	Cases occur randomly through time. ***t*** is the total number of time periods.
H_a	Cases cluster about a single point in time.

Test statistic

The test statistic K, measures the tendency of time periods with at least one case to form a single cluster in time. It is

$$K = \sum_{i=1}^{m} |y_i - y_{r+1}|$$

Where:

- m is the number of time periods with at least one case.
- y_i is the time assigned to the *ith* cell in which a case occurred.
- *(r+1)* is the index of the 'central most' time cell that contained a case, $r = \lceil m/2 \rceil$.
- The floor function of $\lceil x \rceil$ is the largest integer that is less than or equal to x.
- This method measures dispersion of cases about a central time period. K will be small when cases form a single time cluster.

Significance

ClusterSeer calculates the expected test statistic *(E(K))*, the variance *(Var(K))*, z-scores, grand z-scores, lower-tail P-values and overall P-values to help you evaluate the significance of the test statistic.

The expectation and variance of K under the null hypothesis are:

$$E(K) = \frac{(t+1)\left\lceil \frac{m}{2} \right\rceil \left\lceil \frac{m+1}{2} \right\rceil}{m+1}$$

$$Var(K) = \frac{r(t+1)(t-m)((m+1)^2 - 2r^2 - \delta(m))}{12\left(2\left\lceil \frac{m+1}{2} \right\rceil + 1\right)^2}$$

Where $(m) = r - 2$ when m is odd, and $(m) = 2r - 1$ when m is even.

Larsen's method uses the z-score to determine whether occupied time cells within an area tend to occur in a sequence. This method uses an overall z-score to identify unusual pattern over time which may not necessarily be the same over the individual areas.

$$z = \frac{K - E(K)}{\sqrt{Var(K)}}$$, as distributed as N (0,1)

- A z-score of 0 is consistent with a random allocation of occupied cells across the time series.
- *K* will be smaller than *E(K)* when occupied time intervals form a unimodal cluster, and the z-score will be less than 0.
- *K* will be large when occupied time intervals form several clusters. Thus, Larsen's method cannot distinguish a uniform distribution from multiple clusters.
- A uniform distribution of occupied time intervals through time, such as '01010101' will cause *K* to be larger than *E(K)*, and the z-score will be greater than 0.

Significance is therefore evaluated as a one-tailed test describing the probability, under the null hypothesis, of obtaining a *K* as small or smaller than the observed.

The **lower-tail P-value for *K*** is obtained by comparing the z-score to the percentiles of the normal distribution.

When the data consist of several time series, the *K* statistics from each time series can be combined into a **grand z-score** as:

$$Z_G = \frac{\sum_{i=1}^{S} K_i - \sum_{i=1}^{S} E(K_i)}{\sqrt{\sum_{i=1}^{S} Var(K_i)}}$$

This grand z-score tests for an overall departure from the expected values across all time series simultaneously. The individual z-scores test for unimodal clustering within each time series. You must examine the individual z-scores before concluding whether a significant grand z-score is due to unimodal clustering in all of the time series, or to some other combination of temporal pattern across time series.

The overall P-value for *K* across the simultaneous time series is obtained by comparing the grand z-score to the percentiles of the normal distribution.

Note:

- The grand z-score is not biased by differences in population size across time series.
- *K* will also be large when occupied time intervals form several clusters, and Larsen's method thus cannot distinguish a uniform distribution from multiple clusters.

Larsen's Method: How to

Choose Larsen's method from the Analysis menu (Temporal | Single or Several submenus).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data file (text file or DBF).

   If the text file does not include a label, ClusterSeer will use the row number as the label, which assumes that the sequence of case counts in the file increases with the row number. Labels are required with DBFs.

   If you submit a DBF, ClusterSeer will prompt you to identify which columns in your file contain the required data. In this case, the columns can be in any order.

   If you wish to submit a text file without a header, it should contain group-level data with the following columns in the following order:

For single time series:

Time Sequence Label (required for DBF files only)	Case Count

For several time series:

Region Label (required for DBF files only)	Case Count Time 1	Case Count Time 2	Case Count Time 3	...

2. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

3. In the Provide data dialog, you may use the Select File button to change your file choices.

4. Enter the number of time series in your dataset. For a single time series, enter 1. For a several time series, enter the total number of locations where you collected data for consecutive intervals of equal length.

Run the analysis

5. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.

Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Larsen's Method: Results

Plot

Choose Plot from the View menu to view the plot.

ClusterSeer plots the observed statistic *K* on its expectation *E(K)* for both the single and several time series. The time series are the red point or points on the graph. The dashed blue line is the function *K = E(K)* describing where observations would be plotted under the null hypothesis of a random distribution of occupied intervals across the time series. Unimodal clustering will cause *K* to be smaller than its expectation, and observations plot below the dashed line. Examine the results of the significance test for the z-score to determine whether or not there is unimodal clustering of cases.

Session log

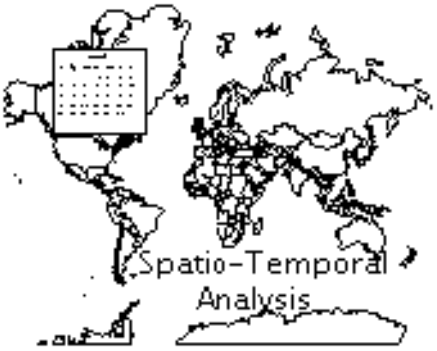
After ClusterSeer performs a Larsen analysis, it will place summary information and results into the session log.

Parameters and summary statistics for each time series:

- The number of the series
- The number of cells per series
- The test statistic, *K*, measures the tendency of time periods with at least one case to form a single cluster in time.
- The expectation of *K*, *E(K)*
- The variance of *K*, *Var(K)*
- Larsen's method uses the z-score to determine whether occupied time cells within an area tend to occur in a sequence.
- The Lower-tailed P-value for *K* is obtained by comparing the z-score to the percentiles of the normal distribution.
- This method uses the Grand z-score, to identify unusual overall pattern over time which may not necessarily be the same over the individual areas.
- For the multiple time series, ClusterSeer reports an Overall P-value for *K* across the simultaneous time series. This value is obtained by comparing the Grand z-score to the percentiles of the normal distribution.

CHAPTER 28

Mantel's Method



Mantel's method (Mantel 1967) quantifies space-time interaction for individual-level data. It does not require specifying critical or threshold distances for space-time association, unlike Knox's method. The method calculates space and time distance matrices. The test statistic, *r*, is the sum of the time distance multiplied by the spatial distance for all case pairs.

Examples

Chenoweth et al. (2002) used Mantel's test to evaluate the phylogeography of the pipefish Urocampus carinirostris. They analyzed the similarity in mitochondrial DNA and the geographical distribution of clades in the group, finding unexpected patterns not predicted by current biogeographic hypotheses. Schmucki et al. (2002) examined the spatio-temporal relationships among hedgerows in three different agricultural landscapes in Quebec (1958-97). They found significant differences among landscapes that they associated with changes in agricultural techniques and management during that period.

Mantel's Method: Statistic

H_o	The times of occurrence of the health events are distributed randomly across the case locations. This is another way of saying the time distances between pairs of cases are independent of the spatial distances between pairs of cases.
H_a	Pairs of cases near in space tend to be near in time.

Test statistic

Mantel's test statistic, *Z*, is the sum, across all case pairs, of the time distance multiplied by the spatial distance. *Z* is also called the **Mantel product**.

$$Z = \sum_{i=1}^{N} \sum_{j=1}^{N} s_{ij} t_{ij}$$

Where

- N is the number of cases
- s_{ij} is the distance between i and j in space, s is the average space distance, and s_s is the standard deviation of s_{ij}
- t_{ij} is the distance between i and j in time, t is the average space distance, and s_t is the standard deviation of t_{ij}.

ClusterSeer uses the standardized version of the Mantel product, *r*. *r* is a measure of matrix correlation with range $-1 < r < 1$. It is easier to interpret than *Z*. Both *r* and *Z* become large when the time distances are linearly dependent on the space distances.

$$r = \frac{1}{(N^2 - N - 1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{s_{ij} - \bar{s}}{s_s} \frac{t_{ij} - \bar{t}}{s_t}$$

Significance

Although Mantel (1967) provides an approximation for the variance of Z under the null hypothesis of no association between space and time, the usual approach is to generate the distribution of r using Monte Carlo simulations, permuting the elements of one of the distance matrices while holding the other constant. This is equivalent to repeatedly scrambling the time observations across the locations, and calculating r each time. This is done repeatedly to generate a distribution of r under the null hypothesis.

Mantel's Method: Transformations

For a contagious disease we expect the small space and time distances to be correlated, but not the large distances. Once you have performed a Mantel analysis, you can look at the correlation in the distances using the Plot.

To correct for pattern expected in contagious disease, Mantel recommended transforming the space-time distances to reduce the effect of large ones.

You can transform the space-time distances by adding a constant (the shift) to all of the distances, and or raising the distance to a power. You can also take the log of the distance.

$$(d + shift)^{power}$$

Here shift is the constant and d is the distance to be transformed. Mantel specifically recommended the reciprocal transformation $1/(d + \text{constant})$ for transforming contagious diseases (power = -1). If you specify a transformation, ClusterSeer will report the transformation you specified in the Session log using D to represent spatial distance and T to represent temporal distance.

The selection of a transformation and constant can be subjective, and the default settings in ClusterSeer are not to use them (shift = 0 and power = 1). The time distances and the temporal distances can each be transformed separately.

Mantel's Method: How to

Choose Mantel's Method from the Analysis menu. (Spatio-Temporal submenu).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

This analysis requires a single file of case event data. Files should follow ClusterSeer general data requirements.

Submit data file

1. ClusterSeer will prompt you to submit the data file (text file, DBF, or point shapefile).

   If you submit a text file with a header, a shapefile, or DBF, ClusterSeer will prompt you to identify which columns in your file contain the required data. In this case, the columns can be in any order.

   If you wish to submit a text file without a header, it should contain individual-level data with the following columns in the following order

Space-Time Case File:

case level (optional unless you are submitting a DBF)	case event x-coordinate (not for shapefile)	case event y-coordinate (not for shapefile)	case event timepoint (user-defined integer)

ClusterSeer expects case event timepoints as user-defined integers. See "Temporal data formats" on page 174.

2. You will need to specify the coordinate system of the data. If the data are in geographic coordinates, you will also need to choose a distance measurement.
3. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

4. In the Provide data dialog, you may use the Select File button to change your file choices.

5. Enter the distance transformations. See the Transformations section for more information.
6. Choose the number of Monte Carlo runs, the number of simulations used to determine statistical significance of the test statistic. The default value is 999.
7. Enter the significance level you wish to use for the test. The significance level is the alpha level, the cutoff for statistical significance. The default value is 0.05.

Run the analysis

8. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.

Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Mantel's Method: Results

Monte Carlo distribution

You can view the Monte Carlo distribution by choosing MC Distribution from the View menu.

This histogram shows the reference distribution generated by randomizing the dataset and recalculating the observed value. The relative position of the observed value of *r* is illustrated with a slim, vertical black line.

Map

Choose Map from the View menu. ClusterSeer will display a map of the cases' spatial distribution.

Q? If you query one of these points, you can view its label, spatial coordinates, and its time of occurrence.

Plot

You can view the plot by choosing Plot from the View menu.

This plot is a scattergram of all the temporal and spatial distances used in the analysis. If you transformed them, it is a graph of the transformed distances, not the raw data.

Session log

After ClusterSeer performs a Mantel analysis, it will place summary information and results into the session log.

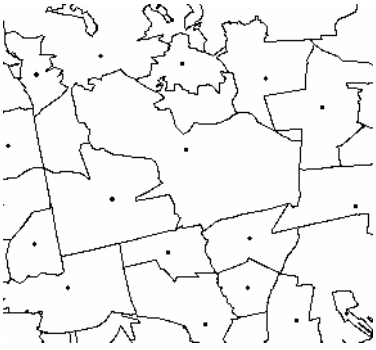
- The file used
- The number of cases analyzed
- The transformations you specified
- Mantel's *r*, the standardized test statistic

Monte Carlo results

- The test statistic *r*
- The number of Monte Carlo simulations
- The P-value for the test statistic through comparison with the Monte Carlo distribution.

CHAPTER 29

Moran's I Method



Moran's *I* (Moran 1950) is a weighted correlation coefficient used to detect departures from spatial randomness. Departures from randomness indicate spatial patterns, such as clusters. The statistic may identify other kinds of pattern such as geographic trend.

Moran's *I* tests for global spatial autocorrelation in group-level data. Positive spatial autocorrelation means that nearby areas have similar rates, indicating global spatial clustering. Nearby areas have similar rates when their populations and exposures are alike. When rates in nearby areas are similar, Moran's *I* will be large and positive. When rates are dissimilar, Moran's *I* will be negative.

Moran's *I* requires full enumeration of the connections among the observations, which may be a problem when the number of areas becomes large. When full enumeration isn't possible, use Grimson's method, and estimate the Grimson input data from a sample of areas. Moran's *I* is biased by large differences in population size across areas. Use Oden's *Ipop* when population size data are available.

Examples

Cullen et al. (2001) used Moran's *I* to examine patterns in silver beech, *Nothofagus menziesii*, population dynamics in New Zealand treelines. They found

significant patchiness in recruitment of silver beech seedlings. Castresana (2002) used Moran's *I* to characterize the pattern of mutation locations on human and mouse chromosomes.

Moran's I Method: Statistic

H_o	Disease rates are spatially independent, the observed rates are assigned at random among locations. ***I*** is close to zero, depending on sample size.
H_a	Disease rates are not spatially independent. ***I*** is not zero.

Test statistic

Moran's *I* (Moran 1950) is used to determine whether neighboring areas are more similar than would be expected under the null hypothesis. Moran's *I* is:

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N, (j \neq i)} w_{ij} z_i z_j}{S_0 \sum_{i=1}^{N} z_i^2}$$

where *N* equals the number of regions, w_{ij} is a weight denoting the strength of the connection between areas *i* and *j*, z_i is the rate in region *i* centered about the mean rate (using $z_i = x_i - \text{ave}(x)$; x_i is the rate in region *i*); and S_o is the sum of the weights.

$$S_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}, i \neq j$$

The expectation of *I* under the null hypothesis is:

$$E(I) = -\frac{-1}{(N-1)}$$

The expectation becomes close to zero as *N* increases. The variance of *I* is determined under two null hypotheses or assumptions: Normality (denoted *N*) or randomization (denoted *R*). Under assumption *N* the rates are sampled from a population whose distribution is normal. Under assumption *R* the rates are random samples from a population whose distribution is unknown. Assumption *N* is useful when we have good reason to believe the observations follow a normal distribution. Assumption *R* is less restrictive and, since we often don't know their theoretical distribution, is appropriate for disease rates. The variance under assumption *N* is:

$$Var_N(I) = \frac{1}{(N-1)(N+1)S_0^2}(N^2S_1 - NS_2 + 3S_0^2) - E(I)^2$$

Under the assumption R the variance is:

$$Var_R(I) = \frac{N[(N^2-3N+3)S_1 - NS_2 + 3{S_0}^2] - b_2[(N^2-N)S_1 - 2NS_2 + 6S_0^2]}{(N-1)^{(3)}S_0^2} - E(I)^2$$

where a falling factorial is written $s^{(b)} = s(s-1)\ldots(s-b+1)$,

and where

- $b_2 = m_4/m_2^2$
- $m_4 = 1/N\sum_{i=1}^{N} z_i^4$

- $m_2 = 1/N \sum_{i=1}^{N} z_i^2$
- $S_1 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1, j \neq 1}^{N} (w_{ij} + w_{ji})^2$
- $S_2 = \sum_{i=1}^{N} (w_{i\bullet} + w_{\bullet i})^2$

Note: For S_2, the positioning of the bullet symbol $(\bullet)$ indicates whether to add columns or rows within the matrix. For instance, $w_{i\bullet}$ symbolizes the sum of elements across rows, and $w_{\bullet i}$ symbolizes the sum of elements within columns.

Moran's I Method: Significance

ClusterSeer evaluates the significance of Moran's *I* under assumptions *R* and *N*, and also by Monte Carlo simulations. See page 20 of the first ClusterSeer user guide for more information on calculating Monte Carlo P-values.

For assumptions *N* and *R* ClusterSeer calculates two z-scores as:

$$z_N = \frac{I - E(I)}{\sqrt{Var_N(I)}} \text{ and } z_R = \frac{I - E(I)}{\sqrt{Var_R(I)}}$$

These z-scores express the difference between the observed and expected value of *I* in standard deviation units. The distribution of the z-scores is approximately normal with a mean of 0 and a variance of 1.0. ClusterSeer reports a two-tailed P-value because spatial pattern is of interest both when Moran's *I* is positive (rates in connected areas are similar) or negative (rates in connected areas are dissimilar).

Moran's I Method: How to

Choose Moran's *I* method from the Analysis menu. (Spatial|Global submenus).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data files (either a shapefile or a text data file and a contiguity file, see p. 43 of the first ClusterSeer User Guide for more information on contiguity files).

   If you submit a shapefile or text file with a header, ClusterSeer will prompt you to identify which columns in your file contain the required data. In this case, the columns can be in any order.

   If you wish to submit a text file without a header, it should contain group-level data with the following columns in the following order:

Centroid label (optional)	Disease rate

2. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

3. In the Provide data dialog, you may use the Select File button to change your file choices.
4. Choose the number of Monte Carlo runs, the number of simulations used to determine statistical significance of the test statistic. The default value is 999.
5. Enter the significance level you wish to use for the test. The significance level is the alpha level, the cutoff for statistical significance. The default value is 0.05.

Run the analysis

6. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.

Then, ClusterSeer will run the Monte Carlo simulation. You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button. Often, the analysis runs so quickly that you will not see the stop button appear.

Moran's I Method: Results

Monte Carlo distribution

You can view the Monte Carlo Distribution by choosing MC Distribution from the View menu.

The histogram shows the reference distribution generated by randomizing the dataset and recalculating the observed value. The relative position of the observed value of Moran's *I* is illustrated with a slim, vertical black line.

Plot

You can view the plot by choosing Plot from the View menu.

The simulated P-value plot shows how the significance of the test statistic changes with the number of Monte Carlo randomizations performed. What you will usually see is that the p-value decreases from near p=1.0 to an asymptote before it reaches the number of randomizations you specified in the analysis.

If it decreases to the asymptote after few randomizations, you specified a greater number of randomizations than was required to fix the P-value.

If it is still jumping around a lot, you may wish to rerun the analysis with a greater number of randomizations to fix the P-value.

Session log

After ClusterSeer performs a Moran's *I* analysis, it will place summary information and results into the session log.

Cluster Seer reports information on the files used including: file name(s), number of regions, total number of regions uniquely identified, and average disease frequency.

Moran's I results

- The value of the test statistic, Moran's *I*
- The expected value of Moran's *I* under the null hypothesis *(E[I])*
- The alpha level at which the test was performed

Normality Assumption results

ClusterSeer reports the variance, the z-score, and the significance of the z-score, obtained from a look-up table.

Randomization Assumption results

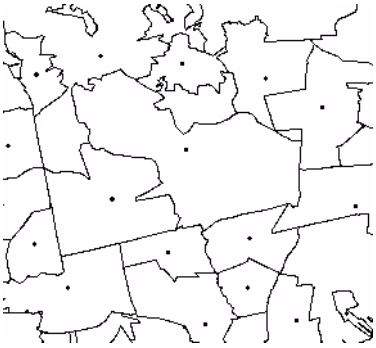
ClusterSeer reports the variance under the randomization assumption, the z-score calculated using this variance, and the significance of the z-score, obtained from a look-up table, S_0 (the sum of the weights), S_1 (A sum used to calculate the variance), S_2 and b_2.

Monte Carlo results

- The value of the test statistic, Moran's *I*
- The number of Monte Carlo simulations
- The two-tailed P-value for the test statistic through comparison with the Monte Carlo distribution.

CHAPTER 30

Oden's Ipop Method



Oden (1995) adapted Moran's *I* to consider population size, developing a new statistic, *Ipop*. *Ipop* is used to detect departures from spatial randomness, but, unlike Moran's *I*, it accounts for differences in population size across areas. If ignored, large differences in population size decrease the ability of Moran's *I* to detect true clustering or departures from spatial randomness.

Like Moran's *I*, *Ipop* explores the global spatial pattern of group-level data.

Example

Fosgate et al. (2002) investigated spatial clustering in human brucellosis in California using the Oden's *Ipop*. They found significant spatial clustering brucellosis cases in several time periods and in several populations.

Oden's Ipop Method: Statistic

H_o	Disease rates in connected areas are independent. The geographic variation in the number of cases is expected to follow geographic variation in population size.
H_a	Disease rates in connected areas are not spatially independent. Geographic variation in the number of cases does not follow geographic variation in population size.

Test Statistic

Moran's *I* (Moran, 1950) is a weighted correlation coefficient used to detect departures from spatial randomness. Moran's *I* is used to determine whether neighboring areas are more similar than would be expected under the null hypothesis. Oden (1995) adjusted Moran's *I* to account for differences in population size across areas. Use *Ipop* when population size data are available.

Ipop requires a fair amount of notation. In essence, *Ipop* is large when there is clustering within a region or among adjacent regions.

$$Ipop = \frac{N^2 \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij}(e_i - d_i)(e_j - d_j) - N(1 - 2\bar{b}) \sum_{i-1}^{m} w_{ij} e_i - N\bar{b} \sum_{i=1}^{m} w_{ii} d_i}{S_0 \bar{b}(1 - \bar{b})}$$

Where:

- m represents the number of locations or areas
- N is the total number of cases in all of the areas
- n_i is the total number of cases in area i
- e_i is the proportion of cases in area i $\left(e_i = \frac{n_i}{N}\right)$

- X is the total size of the risk population in all areas
- x_i is the size of the risk population in area *i*
- d_i is the proportion of the population in area *i*, $d_i = \frac{x_i}{X}$
- $e_i - d_i$ is the difference between the proportion of cases in area *i* and the number of cases expected given the area's population size
- $\bar{b}$ is the average prevalence, $\bar{b} = \frac{N}{X}$; $b_2 = \frac{1}{\bar{b}(1-\bar{b})} - 3$
- $S_0 = X^2 A - XB$
- $S_1 = X^3 E - 4x^2 F + 4XD$
- w_{ij} is a weight denoting the strength of connection between areas i and j, developed from neighbor information.
- $A = \sum_{i=1}^{m} \sum_{j=1}^{m} d_i d_j w_{ij}$
- $B = \sum_{i=1}^{m} d_i w_{ii}$
- $C = \sum_{i=1}^{m} \sum_{j=1}^{m} d_i d_j (w_{ij} + w_{ji})^2$

- $D = \sum_{i=1}^{m} d_i w_{ii}^2$
- $E = \sum_{j=1}^{m} d_i \left[\sum_{j=1}^{m} (w_{ij} + w_{ji}) \right]^2$
- $F = \sum_{j=1}^{m} d_i w_{ii} \sum_{j=1}^{m} d_j (w_{ij} + w_{ji})$
- $G = \sum_{i=1}^{m} e_i w_{ii}$
- $H = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} (e_i - d_i)(e_j + d_j)$

The expectation of *Ipop* under the null hypothesis (no clustering) approaches zero for large total population:

$$E(Ipop) = \frac{-1}{(X-1)}$$

The range of *Ipop* depends on population size, therefore t can be useful to standardize the statistic using the average prevalence, for comparison among different study areas.

$$Ipop' = \frac{Ipop}{\bar{b}}$$

The variance of *Ipop* can be determined based on a random distribution, appropriate for disease rates (Cliff and Ord 1981). ClusterSeer calculates the variance in two ways. The variance of *Ipop* under the null hypothesis is:

$$Var_R(Ipop) = \frac{X[(X^2 - 3X + 3)S_1 - XS_2 + 3S_0^2] - b_2[X^2 S_1 - 2XS_2 + 6{S_0}^2]}{(X-1)^{(3)}{S_0}^2} - E(Ipop)^2$$

It also calculates an approximation of the variance (Var_A):

$$Var_A(Ipop) = \frac{2A^2 + \frac{c}{2} - E}{A^2 X^2}$$

Significance

ClusterSeer evaluates the significance of *Ipop* using several approaches: using the z-scores, variance and multinomial Monte Carlo randomization. In general, these methods will report relatively similar p-values. The approximation and randomization assumption methods are only valid when the data can be assumed to be distributed normally. When the data may not be normally distributed, use the Monte Carlo P-value instead.

Oden's Ipop Method: How to

ClusterSeer requires information on disease frequencies, population-at-risk and neighbor relationships to run Oden's *Ipop*. You can submit this data as two text files: a disease frequency file and an associated contiguity file. Currently, you cannot run an Oden's *Ipop* analysis with a DBF file.

Choose Oden's *Ipop* method from the Analysis menu (Spatial|Global submenus).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data files (either a shapefile or a text data file and a contiguity file, see p. 43 of the first ClusterSeer User Guide for more information on contiguity files)

   If you submit a shapefile or text file with a header, ClusterSeer will prompt you to identify which columns in your file contain the required data. In this case, the columns can be in any order.

   If it is a text file without a header, it should contain group-level data with the following columns in the following order.

Centroid label (optional)	Case count	Population at risk count

2. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

3. In the Provide data dialog, you may use the Select File button to change your file choices.
4. Choose the number of Monte Carlo runs, the number of simulations used to determine statistical significance of the test statistic. The default value is 999.
5. Enter the significance level you wish to use for the test. The significance level is the alpha level, the cutoff for statistical significance. The default value is 0.05.If you run mulitiple tests at the same significance level, you can then choose to run a Multiple Comparisons analysis to determine the proper significance level for all comparisons.

Run the analysis

6. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.

Then, ClusterSeer will run the Monte Carlo simulation.You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Oden's Ipop Method: Results

Monte Carlo distribution

You can view the Monte Carlo distribution by choosing MC Distribution from the View menu.

The histogram shows the reference distribution generated by randomizing the dataset and recalculating the observed value. The relative position of the observed value of *Ipop* is illustrated with a slim, vertical black line.

Plot

You can view the plot by choosing Plot from the View menu.

The simulated P-value plot shows how the significance of the test statistic changes with the number of Monte Carlo randomizations performed. What you will usually see is that the p-value decreases from near p=1.0 to an asymptote before it reaches the number of randomizations you specified in the analysis.

If it decreases to the asymptote after few randomizations, you specified a greater number of randomizations than was required to fix the P-value.

If it is still jumping around a lot, you may wish to rerun the analysis with a greater number of randomizations to fix the P-value.

Session log

After ClusterSeer performs an *Ipop* analysis, it will place summary information and results into the session log.

Information on the files used including the file name(s), number of regions, cases, and population-at-risk.

Oden's Ipop results

- The value of the test statistic, *Ipop*
- The standardized version of *Ipop*, *Ipop'*
- The expected value of *Ipop* under the null hypothesis

- The percent within: the proportion of the test statistic that is attributable to clustering within the area
- The percent among: for adjacent areas, whether those adjacent areas are similar in their level of excess

Approximation

ClusterSeer reports the approximated variance, the z-score, and the significance of the z-score, obtained from a look-up table.

Randomization Assumption results

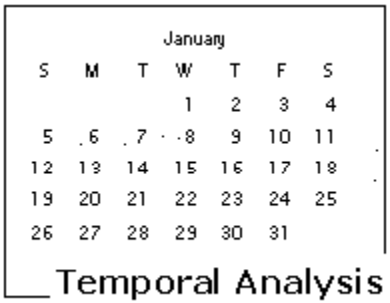
ClusterSeer reports the variance under the randomization assumption, the z-score calculated using this variance, and the significance of the z-score, obtained from a look-up table.

Monte Carlo results

- The test statistic (*Ipop'*)
- The number of Monte Carlo simulations
- The P-value for the test statistic through comparison with the Monte Carlo distribution.

CHAPTER 31

Scan Method



The Scan method (Wallenstein 1980 1987) tests for temporal clustering in single and several time series with group-level data. Use the Scan method with case counts, not rates. The data can be relatively sparse or numerous. The time series must be at least five time cells in length.

You can analyze several time series at once using time series files containing several series. However, the P-values for each series will not be combined to yield an overall P-value. The test is biased by changes in population size through time, which can cause significance even when underlying clustering is absent.

Example

Vredevoe et al. (1999) investigated the temporal distribution of equine granulocytic ehrlichiosis in California. They found seasonal clustering in the cases that parallels the life history of the *Ixodes pacificus* tick, indicating it is the most plausible vector among those considered.

Scan Method: Statistic

H_o	Cases occur at random across the time series.
H_a	Cases cluster in some time periods.

Test statistic

The test statistic S_w is the maximum number of cases appearing in a pre-defined window as it is moves continuously along a time series. This number must be an integer. For example, for this time series

0 **1 2** 0 0 2 0 1

and a window width or size of 2, S_w is 3. The significance of S_w is obtained by numerical approximation.

The test is most sensitive to clustering when the scanning window is the same width as natural clusters in the data. When the cases are clustered, the maximum number of cases in the scanning window will be large.

ClusterSeer calculates expected values, significance values, and the normal approximation variance to help you evaluate the significance of the test statistic.

Scan Method: Significance

Statistical significance arises when many cases occur in one cell or when time cells with many cases fit within the scanning window. S_w is larger than its expectation when cases cluster in a few time cells. S_w is smaller than its expectation when cases are uniformly distributed among the time cells.

When this method was developed, computers were too slow to generate Monte Carlo randomizations. Today, the simulated Monte Carlo P-value is easily computed and appropriate to use in all but a few instances. For example, if you have a very large dataset, you should use the normal approximation, instead. The normal approximation is appropriate if the window size is large relative to the total time-series length, and if there is a sufficient number of cases. To be consistent, ClusterSeer continues to calculate and report the truncated estimate and Wallenstein and Neff P-value.

Options for the expectation under the null hypothesis and P-values are: truncated Wallenstein and Neff (*TE[Sw]*), normal approximation (*NE[Sw]*), and the Monte Carlo simulations (*SE[Sw]*).

TE[S_w]

The Truncated Wallenstein and Neff formula, $TE[S_w]$ is

$$E(S_w) \approx \sum_{j=1}^{N} P(j; w, T, N)$$

Where:

N: Number of cases

T: Number of time periods

w: Window width

S_w: Maximum number of cases observed in w as it is slid along the time series

Wallenstein and Neff (*WN*) P-Value significance: The probability of observing, under the null hypothesis, an S_w given the window width w, T time periods and N cases is approximated by (Wallenstein and Neff, 1987):

$$P(m;w, T, N) \approx (mT/w - N + 1)b(m;N, w/T) + 2\sum_{j=m+1}^{N} b(j;N, w/T)$$

$$b(j;n, p) = \binom{n}{j} p^j (1-p)^{n-j}$$

$\binom{n}{j}$ is a binomial coefficient. $P(j;w, T, N)$ is the probability of obtaining, under the null hypothesis, a scan statistic greater than or equal to j. This probability is one-tailed.

NE[S_w]

NE[S_w] is the normal approximation to Wallenstein and Neff's formula shown in the previous section on Truncated Wallenstein and Neff.

N P-value is the normal approximation P-value. The statistics tend to become asymptotically normal when you use the normal approximation.

ClusterSeer also calculates the normal approximation variance.

SE[S_w]

SE[S_w] is the Monte Carlo simulation estimate. The simulation estimate is obtained by randomly distributing cases into time slots, over and over again for the number of Monte Carlo runs, and then running the Scan statistic on all those examples. The mean of these repetitions is *SE*.

S P-value is the Monte Carlo simulation P-value. The P-value corresponds to the proportion of results more extreme than the result of the given data.

Scan Method: How to

Choose Scan method from the Analysis menu (Temporal | Single or Several submenus).

In a series of dialogs, ClusterSeer will prompt you to submit the file to analyze. If you submitted a suitable dataset in the previous analysis, you will jump directly to the Choose settings step listed below.

Submit data file

1. ClusterSeer will prompt you to submit the data file (either a text file or DBF).

   DBFs must have a column of labels. If you submit a text file without a label column, ClusterSeer will use the row number as the label, which assumes that the sequence of case counts in the file increases with the row number.

   The text or DBF file should contain group-level data with the following columns in the following order:

For single time series:

Time Sequence Label (required for DBF files only)	Case Count

For several time series:

Region Label (required for DBF files only)	Case Count Time 1	Case Count Time 2	Case Count Time 3	...

2. If your data file includes labels, choose Selected data file contains label. If your data have no labels, select Use study row # as label.

Choose settings

3. In the Provide data dialog, you may use the Select File button to change your file choices.
4. Enter the window size for your time series. See the previous Scan Statistic section for details.
5. Enter the total number of time series. If you are running a single time series, specify 1. For multiple time series, specify the total number of time series in your study.

6. Choose the number of Monte Carlo randomizations, the number of simulations used to determine the statistical significance of the Test Statistic.

Run the analysis

7. After you hit OK, ClusterSeer will establish nearest neighbor relationships. If you hit Stop at this point, the procedure will cancel.

Then, ClusterSeer will run the Monte Carlo simulation.You may stop the simulations at any time using the Stop button on the progress bar. The stop button will halt the simulations and the results will be displayed for the number of Monte Carlo runs completed by the time you pressed the button.

Scan Method: Results

Monte Carlo distribution

You can view the Monte Carlo Distribution for both the single and several time series by choosing MC Distribution from the View menu.

The histogram shows the reference distribution generated by randomizing the dataset and recalculating S_w. S_w is illustrated in red, and it is compared with the distribution for estimating the one-sided P-value.

Plot

A graph of the S_w on its expectation is shown for the multiple testing regions. When S_w equals its expectation the test is not significant and a line with a slope of 45 degrees results. When clustering exists S_w is large relative to the expected value and the points will be above the 45 degree line.

Under uniformity (e.g. an equal number of cases in each time periods), S_w will be small relative to its expectation and the points will be below the 45 degree line.

Session log

After ClusterSeer performs a Scan analysis, it will place summary information and results into the session log.

Data and analysis input for each time series

- *N*: the total number of cases in the time series
- S_w: the Scan statistic is the maximum number of cases appearing in a pre-defined window as it moves continuously along a time series.

Results

- The truncated Wallenstein and Neff expectation of the scan statistic (*TE[Sw]*)
- The normal approximation to the expectation of the scan statistic (*NE[Sw]*)
- The normal approximation variance, which measures the dispersion of a distribution about its mean value

- The simulated estimate of expectation of scan statistic ($SE[Sw]$): ClusterSeer performs a Monte Carlo randomization of the data by shuffling the labels for each of the spatial locations.

P-values

- Wallenstein and Neff P-value: the probability of observing, under the null hypothesis, an S_w given the window width w; T time periods and N cases.
- Normal P-value is the normal approximation P-value: the statistics tend to become asymptotically normal when you use the normal approximation.
- ClusterSeer generates simulation P-values for the Monte Carlo randomization for each time series. This is a way to compare the observed S_w to the distribution of S_w based on a random distribution of the data.

ClusterSeer also reports a list of time series it was unable to analyze.

References

Anselin, L. 1995. Local indicators of spatial association-LISA. Geographical Analysis 27: 93-115.

Bailey, T.C., and A.C. Gatrell. 1995. Interactive Spatial Data Analysis. Harlow, UK: Longman Scientific & Technical.

Barbujani, G., and E. Calzolari. 1984. Comparison of two statistical techniques for the surveillance of birth defects through a Monte Carlo simulation. Statistics in Medicine 3: 239-47.

Bender, A.P., A.N. Williams, R.A. Johnson, and H.G. Jagger. 1990. Appropriate public health responses to clusters: the art of being responsibly responsive. American Journal of Epidemiology 132: S48-S52.

Besag, J., and J. Newell. 1991. The detection of clusters in rare diseases. Journal of the Royal Statistical Society Series A 154: 143-155.

Bithell, J.F. 1995. The choice of test for detecting raised disease risk near a point source. Statistics in Medicine 14: 2309-2322.

Bithell, J.F. 1999. Disease mapping using the relative risk function estimated from areal data. Disease Mapping and Risk Assessment for Public Health. A.B. Lawson, A. Biggeri, D. Bohning, E. Lesaffre, J.-F. Viel, and R. Bertollini, eds. New York: John Wiley & Sons. pp. 247-55.

Bithell, J.F., S.J. Dutton, N.M. Draper, and N.M. Neary. 1994. Distribution of childhood leukemias and non-Hodgkin's lymphomas near nuclear installations in England and Wales. British Medical Journal 309: 501-505.

Burra, T., M. Jerrett, R.T. Burnett, and M. Anderson. 2002. Conceptual and practical issues in the detection of local disease clusters: a study of mortality in Hamilton, Ontario. Canadian Geographer 46: 160-71.

Caldwell, G.G. 1990. Twenty-two years of cancer cluster investigations at the Centers for Disease Control. American Journal of Epidemiology 132: S43-47.

Castresana, J. 2002. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. Nucleic Acids Research 30: 1751-6.

Ceccato, V., and L.O. Persson. 2002. Dynamics of rural areas: an assessment of clusters of employment in Sweden. Journal of Rural Studies 18: 49-63.

Centers for Disease Control. 1990. Guidelines for investigating clusters of health events. Mortality and Morbidity Weekly Report 39: 1-16.

Chenoweth, S.F., J.M. Hughes, and R.C. Connolly. 2002. Phylogeography of the pipefish, *Urocampus carinirostris*, suggests secondary intergradation of ancient lineages. Marine Biology 141: 541-7.

Cliff, A.D., and J.D. Ord. 1981. Spatial Processes, Models and Applications. Pion, London.

Cullen, L.E., G.H. Stewart, R.P. Duncan, and J.G. Palmer. 2001. Disturbance and climate warming influences on New Zealand *Nothofagus* tree-line population dynamics. Journal of Ecology 89: 1061-71.

Cuzick, J., and R. Edwards. 1990. Spatial clustering for inhomogeneous populations. Journal of the Royal Statistical Society Series B 52: 73-104.

Dat, N.V. 1982. Tests for Time-Space clustering of Disease. Ph. D. dissertation, Dept. of Biostatistics, SPH, University of North Carolina, Chapel Hill, NC.

Diggle, P.J. and B.S. Rowlinson. 1994. A conditional approach to point process modeling of elevated risk. Journal of the Royal Statistical Society 157: 433-440.

Diggle, P.J. 1990. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. Journal of the Royal Statistical Society 153: 349-362.

Dockerty, J.D., K.J. Sharples, and B. Borman. 1999. An assessment of spatial clustering of leukaemias and lymphomas among young people in New Zealand. Journal of Epidemiology and Community Health 53: 154-8.

Doherr, M.G., T.E. Carpenter, W.D. Wilson, and I.A. Gardner. 1999. Evaluation of temporal and spatial clustering of horses with *Corynebacterium pseudotuburculosis* infection. American Journal of Veterinary Research 60: 284-91.

Doherr, M.G., A.R. Hett, J. Rufenacht, Z. Zurbriggen, and D. Heim. 2002. Geographical clustering of cases of bovine spongiform encephalopathy (BSE) born in Switzerland after the feed ban. Veterinary Record 151: 467-72.

Ederer, F., M.H. Myers, and N. Mantel. 1964. A statistical problem in space and time: Do leukemia cases come in clusters? Biometrics 20: 626-638.

Fishman, G.S. 1973. Concepts and Methods in Discrete Event Digital Simulation. New York: John Wiley and Sons.

Fosgate, G.T., T.E. Carpenter, B.B. Chomel, J.T. Case, E.E. DeBess, K.F. Reilly. 2002. Time-space clustering of human brucellosis, California, 1973-1992. Emerging Infectious Diseases 8: 672-8.

Getis, Arthur and J.K. Ord. 1992. The analysis of spatial association by use of distance statistics. Geographical Analysis 24: 189-206.

Gilman, E.A., R.J.Q. McNally, and R.A. Cartwright. 1999. Space-time clustering of acute lymphoblastic leukaemia in parts of the UK (1984-1993). European Journal of Cancer 35: 91-96.

Grau, H.R. 2002. Scale-dependent relationships between treefalls and species richness in a neotropical montane forest. Ecology 83: 2591-2601.

Grimson, R. 1993. Disease clusters, exact distributions of maxima and p-values. Statistics in Medicine 12: 1773-94.

Grimson, R.C., and R.D. Rose. 1991. A versatile test for clustering and a proximity analysis of neurons. Methods of Information in Medicine 30: 299-303.

Grimson, R.C. 1989. Assessing patterns of epidemiologic events in space-time. In Proceedings of the 1989 Public Health Conference on Records and Statistics. Hyattsville, MD: National Center for Health Statistics.

Hjalmars, U., M. Kulldorff, G. Gustafsson, and N. Nagarwalla. 1996. Childhood leukemia in Sweden: Using GIS and a spatial scan statistic for cluster detection. Statistics in Medicine 15: 707-175.

Holland, B.S. and M.D. Copenhaver. 1987. An improved sequentially rejective Bonferroni test procedure. Biometrics 43: 417-23.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6: 65-70.

Jacquez, G.M. 1996. A *k*-nearest neighbor test for space-time interaction. Statistics in Medicine 15: 1935-49.

Jacquez, G.M. 1994. User manual for Stat! Statistical software for the clustering of health events. Ann Arbor, MI: BioMedware.

Jacquez, G.M. and D.A. Greiling. 2002. The geographic distribution of breast, lung and colorectal on Long Island, New York. In review. www.biomedware.com

Jacquez, G. M. and L.A. Waller. 1999. The effect of uncertain locations on disease cluster statistics. In Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and

Remote Sensing. H. T. Mowrer and R. G. Congalton, eds. Chelsea, Michigan: Sleeping Bear Press. pp 53-64.

Jeffery, J.A., P.A. Ryan, S.A. Lyons, P.T. Thomas, and B.H. Kay. 2002. Spatial distribution of vectors of Ross River Virus and Barmah Forest virus on Russell Island, Moreton Bay, Queensland. Australian Journal of Entomology 41: 329-38.

Knox, G. 1964. The detection of space-time interactions. Applied Statistics 13: 25-29.

Knox, G. 1963. Detection of low intensity epidemicity: application to cleft lip and palate. British Journal of Preventive and Social Medicine 18: 17-24.

Kulldorff, M. 1999. Spatial scan statistics: models, calculations, and applications, in Scan Statistics and Applications. J. Glaz and N. Balakrishnan, eds. Boston: Birkhauser. pp. 303-322.

Kulldorff, M. 1997. A spatial scan statistic. Communications in Statistics Theory and methods 26: 1481-96.

Kulldorff, M., and N. Nagarwalla. 1995. Spatial disease clusters: detection and inference. Statistics in Medicine 14: 799-810.

Kulldorff, M., E.J. Feuer, B.A. Miller, and L.S. Freedman. 1997. Breast cancer clusters in Northeastern United States: a geographic analysis. American Journal of Epidemiology 146: 161-70.

Larsen, R.J., C.L. Holmes and C. W. Heath. 1973. A statistical test for measuring unimodal clustering: a description of the test and of its application to cases of acute leukemia in metropolitan Atlanta, Georgia. Biometrics 29: 301-309.

Lawson, A.B. 1989. Score tests for detection of spatial trend in morbidity data. Dundee: Dundee Institute of Technology.

Le, N.D., A.J. Petkau, and R. Rosychuk. 1996. Surveillance of clustering near point sources. Statistics in Medicine 15: 727-740.

Levin, B. & J. Kline. 1985. The cusum test of homogeneity with an application in spontaneous abortion epidemiology. Statistics in Medicine, 4: 469-488.

Machado-Coelho, G.L.L., R. Assuncao, W. Mayrink, and W.T. Caiaffa. 1999. American cutaneous leishmaniasis in southeast Brazil: space-time clustering. International Journal of Epidemiology 28: 982-9.

Manly, B.F.J. 1986. Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. Researches on Population Ecology 28: 201-218.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27: 209-220.

Moran, P.A.P. 1950. Notes on continuous stochastic phenomena. Biometrika 37: 17-23.

Morganstern, H. 1998. Chapter 23: Ecologic studies. In Modern Epidemiology, 2nd edition. K.J. Rothman and S. Greenland, eds. Philadelphia: Lippincott-Raven. pp.459-80.

Norstrom, M., D.U. Pfeiffer, and J. Jarp. 2000. A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds. Preventative Veterinary Medicine 47: 107-19.

O'Brien, S.J., and P. Christie. 1997. Do CuSums have a role in routine communicable disease surveillance?, Public Health 111: 255-8.

Oden, N. 1995. Adjusting Moran's I for population density. Statistics in Medicine 14: 17-26.

Ord, J.K. and A. Getis. 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis 27: 286-306.

Pagano, J.S. 1999. Epstein-Barr virus: the first human tumor virus and its role in cancer. Proceedings of the Association of American Physicians 111: 573-580.

Page, E.S. 1961. Cumulative sum charts. Techonometrics 3: 1-9.

Page, E.S. 1954. Continuous inspection schemes. Biometrika 41: 100-15.

Ratcliffe, J.H., and M.J. McCullagh. 2001. Chasing ghosts? Police perception of high crime areas. British Journal of Criminology 41: 330-41.

Ripley, B.D. 1976. The second-order analysis of stationary point processes. Journal of Applied Probability 13: 255-66.

Ripley, B.D. 1981. Spatial Statistics. New York: John Wiley & Sons.

Robinson, D. and J.D. Williamson. 1974. Cusum charts. The Lancet i: 317.

Rogerson, P.A. 1997. Surveillance systems for monitoring the development of spatial patterns. Statistics in Medicine 16: 2081-2093.

Rothman, K.J. and S. Greenland. 1998. Measures of disease frequency & measures of effect and measures of association. In: Modern Epidemiology. Philadelphia: Lippincott-Raven. pp. 29-64.

Sarkar, S.K., and C.-K. Chang. 1997. The Simes method for multiple hypothesis testing with positively dependent test statistics. Journal of the American Statistical Association 92: 1601-8.

Schulte, P.A., R.L. Ehrenberg, and M. Singal. 1987. Investigation of occupational cancer clusters: theory and practice, American Journal of Public Health 77: 52-6.

Schmucki, R., S. DeBlois, A. Bouchard, and G. Domon. 2002. Spatial and temporal dynamics of hedgerows in three agricultural landscapes of southern Quebec, Canada. Environmental Management 30: 651-664.

Simes, R.J. 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 751-4.

Snow, J. 1855. On the Mode of Communication of Cholera. London: John Churchill.

Sokal, R.R., N.L. Oden, & B.A. Thomson. 1988. Local spatial autocorrelation in a biological model. Geographical Analysis 30: 331-354.

Tango, T. 1995. A class of tests for detecting "general" and "focused" clustering of rare diseases. Statistics in Medicine 14: 2323-2334.

Turnbull, B.W., E.J. Iwano, W.S. Burnett, H.L. Howe, and L.C. Clark. 1990. Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. American Journal of Epidemiology 132: S136-S143.

van Buuren, S., B.M. Zaadstra, C.P. Zwanikken, D. Buljevac, and J.M. van Noort. 1998. Space-time clustering of multiple sclerosis cases around birth. Acta Neurologica Scandinavica 97: 351-8.

Vredevoe, L.K., P.J. Righter, J.E. Madigan, and R.B. Kimsey. 1999. Association of *Ixodes pacificus* (Acari: Ixodidae) with the spatial and temporal distribution of equine granulocytic ehrlichiosis in California. Journal of Medical Entomology 36: 551-561

Wallenstein, S. and N. Neff. 1987. An approximation for the distribution of the scan statistic. Statistics in Medicine 6: 197-207.

Wallenstein, S. 1980. A test for detection of clustering over time. American Journal of Epidemiology 104: 576-584.

Waller, L.A., and G.M. Jacquez. 1995. Disease models implicit in statistical tests of disease clustering. Epidemiology 6: 584-90.

Waller, L.A., and B.W. Turnbull. 1994, The effect of scale on tests of disease clustering. Statistics in Medicine 12: 1969-84.

Waller, L.A., B.W. Turnbull, L.C. Clark, and P. Nasca. 1994. Spatial pattern analyses to detect rare disease clusters. In Case Studies in Biometry. N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhourse eds. New York: John Wiley & Sons. pp. 13-16.

Waller, L.A., B.W. Turnbull, L.C. Clark, and P. Nasca. 1992. Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and TCE-contaminated dumpsites in upstate New York. Environmetrics 3: 281-300.

Ward, M.P. and T.E. Carpenter. 2000. Techniques for analysis of disease clustering in space and time in veterinary epidemiology. Preventive Veterinary Medicine 45: 257-84.

Williams, E.H., P.G. Smith, N.E. Day, A. Geser, J. Ellice, and P. Tukei. 1978. Space-time clustering of Burkitt's lymphoma in the West Nile District of Uganda. British Journal of Cancer 37: 109-122.

Glossary

A

alpha level: Synonym for significance level, a probability threshold used for evaluating a null hypothesis.

alpha parameter: A parameter used to determine the shape of the raised density function in Diggle's method.

alternative hypothesis: An alternative to the null hypothesis, a different prediction defined either in terms of the null spatial model or in terms of additional parameters to define "clustering".

alternative spatial model: An alternative to the null spatial model. It can be very basic, "not the null spatial model," or it can be a more specific model defining a particular disease distribution.

autocorrelation (spatial): Positive spatial autocorrelation means that nearby areas have similar rates, indicating spatial clustering. Nearby areas have similar rates when their populations and exposures are alike.

average disease frequency: Disease frequency estimated from the dataset itself, the ratio of the total case count over the total population at risk.

B

baseline disease frequency: Used as a reference to evaluate suspected change in disease frequency. A national or historic frequency may be used as the expected frequency or it may be estimated as the average frequency calculated for the study population under investigation.

C

calendar-based intervals: Any method for recording times for temporal data that is based on the calendar year, such as daily, weekly, monthly, or yearly intervals. User-defined data is not directly based on the calendar.

case: A study subject that has experienced a health-related event (usually identified as disease diagnosis). Case data may catalog individuals, or cases may be aggregated into groups for disease frequency or case count data.

case-control status: Indicated with a 1 (integer) if subject is a case and 0 if subject is a control

case count: The number of cases in a particular location, at a particular time, or both.

census data: Information from surveys of population size reported for various years. Within ClusterSeer, census data can be used to estimate population-at-risk size.

cluster: An aggregation of disease in space, in time, or in both space and time, often considered the same as a 'disease outbreak'

contiguity relationship: Continuity, or the state of being so near as to be touching. Within ClusterSeer, two regions are defined as contiguous if they share a common border. See rook and/or queen.

control: A study subject that has not experienced the health-related event under investigation.These subjects are considered to represent all individuals at risk of illness and are used for comparison purposes to uncover factors that may influence risk of disease.

coordinate system: A method for representing spatial location. Within ClusterSeer, spatial information can be represented using any planar projection and geographic coordinates, though geographic coordinates are transformed to UTM for analysis.

D

data format: Within ClusterSeer, data format refers to the data import requirements for different types of data.

data type: Within ClusterSeer, data type refers to the unit of observation in the dataset: whether it describes individuals or groups.

dataset: The observations used for analysis. The dataset for a particular method may be found in one or several files.

disease frequency: Measurement of a change in health status (disease state); usually calculated as an incidence proportion by dividing the case count by the population-at-risk count. It may be calculated locally (over time or over space) for comparison to either the average or expected disease frequency.

E

e: E stands for "exponent" in scientific notation. For example, 3e-005= (3.0 X 10 raised to the negative 5, or 0.00003). Note that ClusterSeer uses as many zeros as necessary as placeholders for exponential values.

ego: A target region, in defining spatial weight files.

expected disease frequency: A disease frequency value supplied by you when specifying a ClusterSeer method. It is usually estimated from another population, for comparison with the study data.

extrapolation: A set of processes for estimating values in between and outside of samples. Within ClusterSeer, you may extrapolate census data with linear or step methods.

F

focus: Point location of potential environmental exposure. ClusterSeer offers methods for evaluating the pattern of disease relative to a focus.

G

global clustering: As used within ClusterSeer and its help, global clustering methods are tests that evaluate clustering by looking at spatial patterns throughout the entire study area. Contrast with local or focused methods.

group-level data: A data type where units of observation are collections of study subjects aggregated over geographic regions and/or temporal intervals. Compare to individual-level data.

I

individual-level data: A data type where the units of observation are subjects that are cases or controls. Compare to group-level data.

inhomogeneous: Not uniform.

intensity: Determines the expected number of points or cases per unit area for Poisson point process null models.

interquartile distance: The difference between the values for the 25th-percentile and the 75th-percentile of a distribution. Used in the local Moran method.

L

label: When importing data, labels are used to match data imported in separate files. The term can also refer to editable text labels on the axes of histograms and plots.

local clustering: As used within ClusterSeer and its help, local clustering methods are tests that evaluate clustering by looking at the level of individual cases or regions within the study area. Contrast with global or focused methods.

M

Monte Carlo Randomization (MCR): A computationally intense method that estimates probability values through resampling the dataset. MCR involves repeatedly reassigning observations to sample locations in a random way, according to a particular null hypothesis, and recalculating the statistic for the sets of randomized data.

N

nested: A polygon completely contained within another polygon, a nested polygon only shares borders with the polygon that contains it.

null distribution: The distribution of the test statistic based on the null hypothesis. It can be derived empirically through Monte Carlo randomization or through distribution theory.

null hypothesis: A prediction based on the null spatial model.

null spatial model: Defines the distribution of cases of the disease expected without clustering.

O

one-tailed P-value: A P-value obtained by comparing the test statistic to one end of the reference distribution. Most ClusterSeer methods are one-tailed, focusing on the upper tail. They test for clustering, for where test statistics will be higher than expected.

P

P-value: The probability that the observed test statistic was drawn from the null distribution, or the probability that the null hypothesis is true given the observed statistic.

point data: Data from individual spatial locations. Points may represent the locations of individual disease cases, or they may represent region centroids for group-level data.

polygon data: Data representing regions as areas.

polygon, nested: A polygon completely contained within another polygon, a nested polygon only shares borders with the polygon that contains it.

polygon, self-intersecting: A polygon is called "self-intersecting" when two or more of its borders intersect anywhere except their endpoints. Make sure to prepare your data with a GIS data editor so that it contains no self-intersecting polygons.

population-at-risk: The individuals considered at risk for the health event (i.e. disease) under investigation. This value serves as a reference population during cluster analysis. Populations-at-risk may also be divided into subpopulations (i.e. based on location or age) and these subpopulation counts can serve-as or contribute-to the units of analysis. If a disease is rare, the cases may be included in the population-at-risk as would be expected with census data.

Q

queen contiguity: Two regions are defined as contiguous under the queen criterion if they share a border of any length, even a single point such as a corner. Compare to rook.

R

reference distribution: A distribution of the test statistic under the null hypothesis, usually obtained by Monte Carlo simulations or from distribution theory.

region: Within ClusterSeer and its help file, the term region is used to indicate an area represented by aggregate data. While a region may be outlined with borders, its data is often assigned to its centroid.

region centroid: A point that informally represents a sample area, used for data aggregated within geographic regions. The observations from that region (such as case count, population at risk count) are located to the centroid. Within ClusterSeer centroids are used to establish inter-region distances.

relative risk: The proportional change in risk after exposure, the risk after exposure divided by the baseline risk.

risk: The average probability of disease developing in an individual during a specified time interval.

rook contiguity: Two regions are defined as contiguous under the rook criterion if they share a border of any length greater than a single point. Compare to queen.

S

self-intersecting: A polygon is called "self-intersecting" when two or more of its borders intersect anywhere except their endpoints. Make sure to prepare your data with a GIS data editor so that it contains no self-intersecting polygons.

significance level: A probability threshold used for evaluating a null hypothesis.

spatial weights matrix: A way to represent contiguity relationships between study regions. Each matrix element corresponds to the relationship for a pair of regions.

study area: The entire geographic extent of the data. The study area may be subdivided into regions, represented by aggregate data. Alternatively, the data may describe spatial locations for individual data.

susceptible: Individuals who could contract the studied disease. These individuals may be included in an analysis as the population-at-risk or controls.

T

test statistic: A value summarizing an aspect of the data.

U

upper-tail P-value: A P-value obtained by comparing the test statistic to the end of the reference distribution where the statistic's values are highest. Most ClusterSeer methods are one-tailed, focusing on the upper tail. They test for clustering, where test statistics will be higher than expected.

W

weight: A value used to alter the influence of another variable. Within ClusterSeer, weights are used for edge correction in Ripley's K-function, to specify neighbor relationships for Local Moran and Moran's I, and to include distance from a focus in Lawson and Waller's Score or between neighboring regions in Rogerson's Spatial Pattern Statistic.

Z

z-score: A method of standardization that involves subtracting the expected value (i.e., mean) and dividing by the standard deviation. Z-scores can be interpreted as the number of standard deviation units from the expected value.

Index