Clustering II

Lecture #17 | GEOG 510
GIS & Spatial Analysis in Public Health

Varun Goel

Outline

- Local cluster detection
  - Autocorrelation-based
  - Kulldorff's Scan Statistic (SaTScan)
  - Density-based clusters

Clustering?

- What is clustering?
  - Clustering
    - Identifying whether events/values are clustered in space
    - Global, does not tell us "where"
      - Pattern is not random
    - For observations with values, spatial autocorrelation
    - For unmarked points, other tests

Clustering?

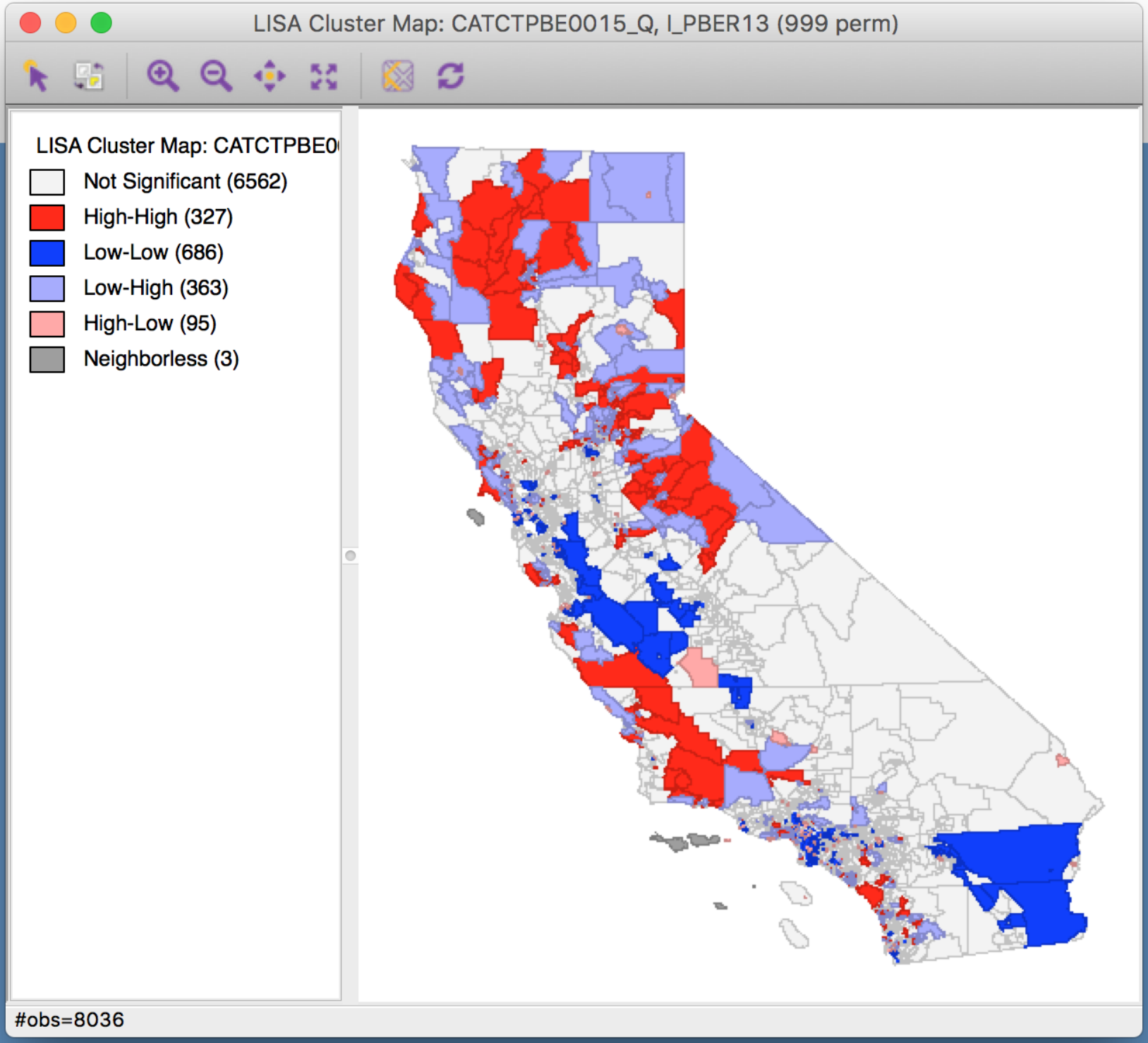
- What is clustering?
  - Cluster detection
    - Identifying clusters of events/values in space (deviations from expected)
    - Local regions having...
      - Higher density (unmarked points)
      - High/Low values (e.g., incidence rate)
    - Spatial autocorrelation, scan statistics

Cluster Detection

- Using Local Autocorrelation
  - For detection of clusters of "high" values, such as elevated incidence rate of some disease, e.g., LISA
  - Simply identifies observations with high values...
    - ...w/ neighbors having high values (H-H)
    - ...w/ neighbors having low values (outlier)

Cluster Detection

- Using Local Autocorrelation
  - LISA method
    - Observation + neighbors
  - LISA results
    - Observation-based
  - Regional clusters are simply a collection of individual observations that happen to be located near one another
    - Nothing tying multiple regions together


LISA Cluster Map: CATCTPBE0015_Q, I_PBER13 (999 perm)
LISA Cluster Map: CATCTPBE0
Not Significant (6562)
High-High (327)
Low-Low (686)
Low-High (363)
High-Low (95)
Neighborless (3)
#obs=8036

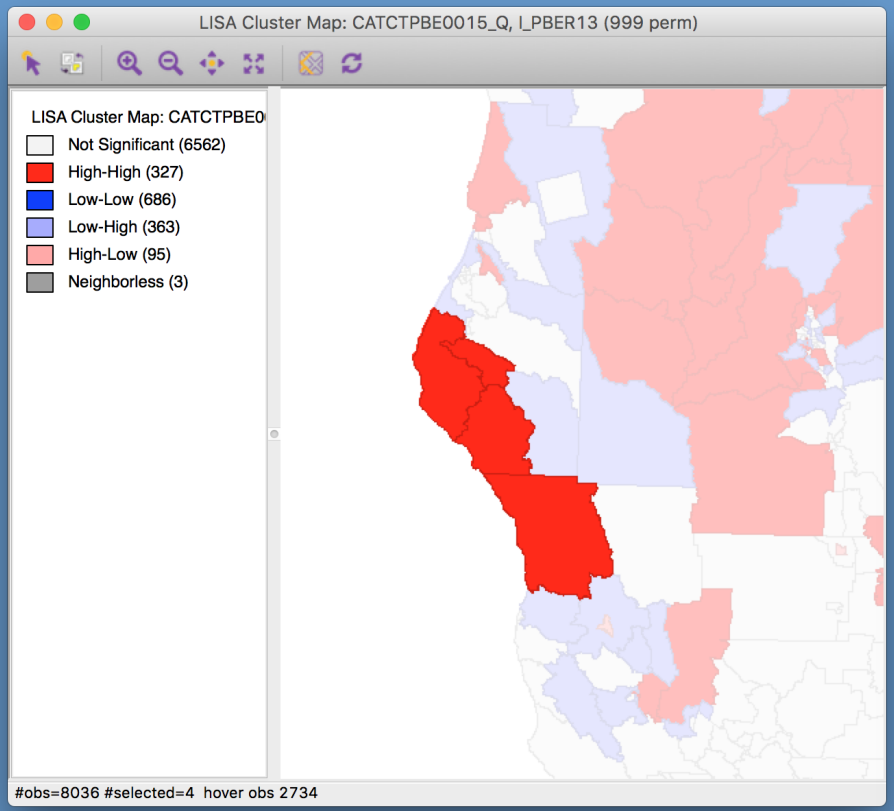


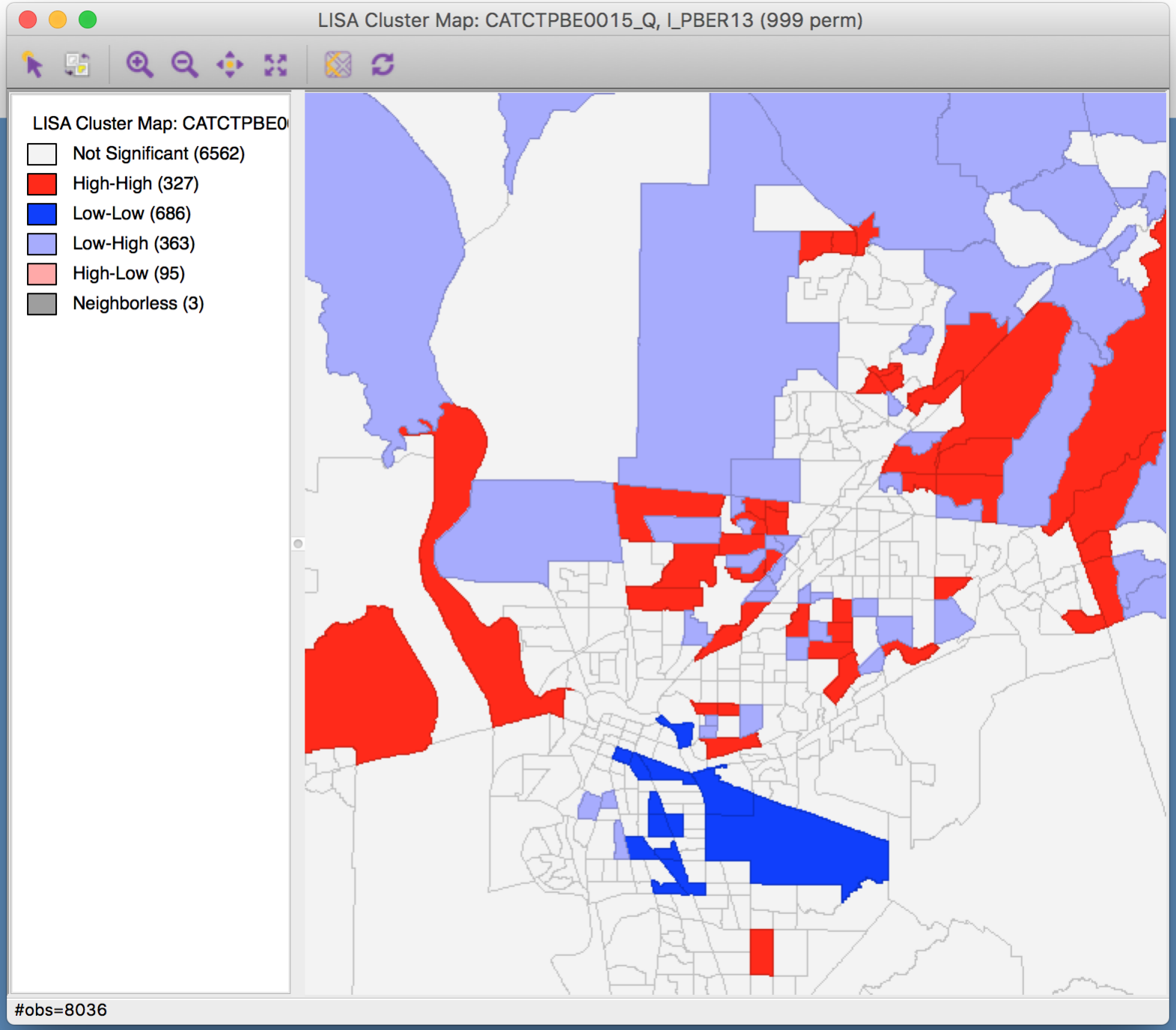
Table - CATCTPBE0015

	R12	PBER13	PBER14	PBER15	LISA_I	LISA_CL	LISA_P
498	15789	0.600000	0.290909	0.117647	27.7429107	1	0.0060000
2827	57692	0.049180	0.048387	0.084746	1.8241487	1	0.0010000
2833	21101	0.370787	0.406250	0.369369	24.6830557	1	0.0050000
3487	95238	0.195122	0.225806	0.066667	5.6619757	1	0.0110000
1	00000	0.059524	0.013889	0.013333	0.1447826	0	0.2190000
2	00000	0.008621	0.050000	0.010000	-0.0776767	0	0.1390000
3	18868	0.000000	0.008929	0.000000	0.1503457	0	0.1870000
4	00000	0.000000	0.000000	0.000000	0.1540371	0	0.1210000
5	20408	0.031746	0.000000	0.018692	-0.0700253	0	0.0800000
6	18018	0.009434	0.000000	0.000000	0.0949752	2	0.0270000
7	00000	0.045455	0.000000	0.000000	-0.1869132	4	0.0010000
8	00000	0.007692	0.008403	0.000000	-0.1011829	0	0.1330000
9	00000	0.000000	0.000000	0.000000	-0.1869093	0	0.1120000
10	00000	0.000000	0.000000	0.000000	0.1688026	2	0.0400000
11	09709	0.000000	0.000000	0.000000	0.1596978	2	0.0140000
12	00000	0.000000	0.000000	0.000000	0.1176653	0	0.1540000
13	00000	0.000000	0.583333	0.684211	-0.2310434	0	0.0740000
14	35714	0.019608	0.021277	0.019802	0.0038928	0	0.4210000
15	16667	0.035714	0.012346	0.013889	-0.0360511	0	0.4920000
16	08547	0.000000	0.000000	0.035211	0.1440382	0	0.1110000
17	99338	0.091837	0.056818	0.023585	0.2728499	0	0.2180000
18	00000	0.000000	0.000000	0.000000	-0.1822374	0	0.1030000
19	00000	0.000000	0.000000	0.000000	0.1470644	0	0.1300000
20	00000	0.000000	0.014286	0.000000	0.1577026	0	0.0930000
21	12658	0.007194	0.005618	0.000000	0.0603476	0	0.3650000
22	03774	0.000000	0.000000	0.002976	0.0792863	0	0.4070000

#obs=8036 #selected=4

Cluster Detection

- Using Local Autocorrelation
  - Regional clusters are simply a collection of individual observations that happen to be located near one another
    - Nothing tying multiple regions together
  - Interpretation can be difficult in some cases


LISA Cluster Map: CATCTPBE0015_Q, I_PBER13 (999 perm)
LISA Cluster Map: CATCTPBE0
Not Significant (6562)
High-High (327)
Low-Low (686)
Low-High (363)
High-Low (95)
Neighborless (3)
#obs=8036

Cluster Detection

- Using Local Autocorrelation
  - Remember, these are rate-based (or proportion) methods
    - No consideration of the size of underlying population in the local cluster
      - Statistical significance of tests are based on number of observation units, not the size of the underlying populations

Cluster Detection

- Using Local Autocorrelation
  - Local cluster size is based on the definition of neighbors
    - Can be small or large
  - But, output remains at the observation-level
    - ...even though the method includes multiple observations in determining that value

Cluster Detection

- Alternate method for areal (polygon) data with rate-based data
  - AMOEBA
    - Space or spatiotemporal
    - Uses Getis Ord Gi*
    - Identifies irregularly shaped clusters
      - Contiguity based

SaTScan

- Kulldorff's Scan Statistic
  - Temporal, Spatial, or Spatiotemporal local cluster detection
    - Considers variation in background population
      - Density/size and attributes
    - Varies observation window size
      - In both space and time, as necessary
    - Compares expected cases to observed cases

SaTScan

- Kulldorff's Scan Statistic
  - Temporal, Spatial, or Spatiotemporal local cluster detection
  - Output unit is a "cluster"
    - A set of observations that belong to the cluster object
      - With associated attributes
    - Output can be multiple clusters
  - *Different from local spatial autocorrelation!*

SaTScan

- Spatial model inputs (cases only)
  - Location coordinates
    - Each location has case and population count
      - These are point representations
      - Thus, for areal (polygon) data, they need to be converted to points (centroids)
  - Prediction coordinates
    - Grid of locations
      - If no grid, locations are used

SaTScan

- Spatial model (cases only)
  - Moving window method that works through the study area
    - Compares observed number of cases to expected number of cases (Poisson) within the scan window
      - Window size varies to account for variable cluster sizes
      - Window shape can be circular, elliptical, or user specified

SaTScan

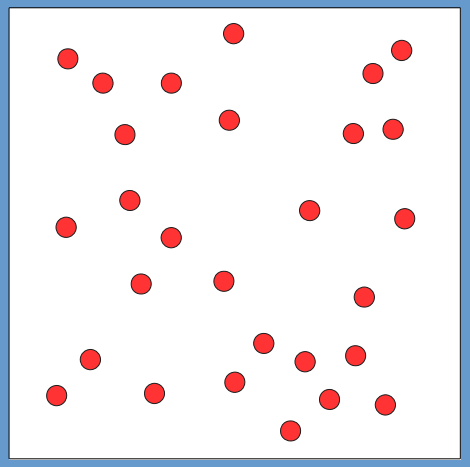
- Spatial model inputs (cases/controls)
  - Location coordinates
    - Cases (1)
    - Controls (0)
      - These are point representations
      - Thus, for areal (polygon) data, they need to be converted to points (centroids)
  - Prediction coordinates
    - Grid of locations
      - If no grid, locations are used

SaTScan

- Spatial model (cases/controls)
  - Moving window method that works through the study area
    - Compares observed number of cases to expected number of cases (Bernoilli) within the scan window
      - Window size varies to account for variable cluster sizes
      - Window shape can be circular, elliptical, or user specified

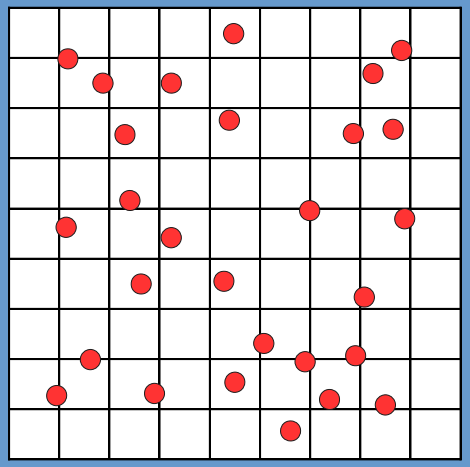
SaTScan

- Moving window method
  - Location of window (grid)



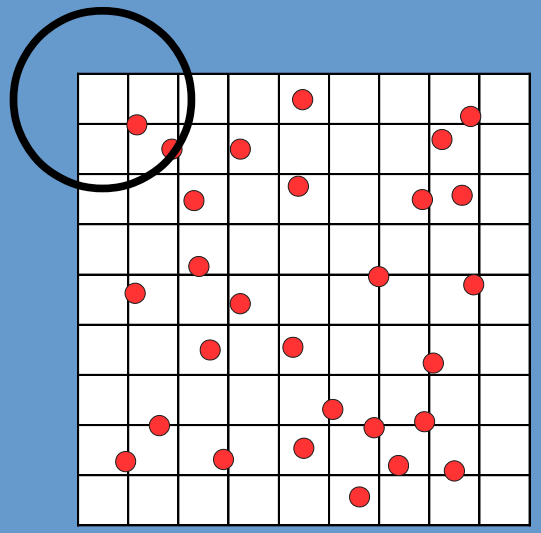
SaTScan

- Moving window method
  - Location of window (grid)



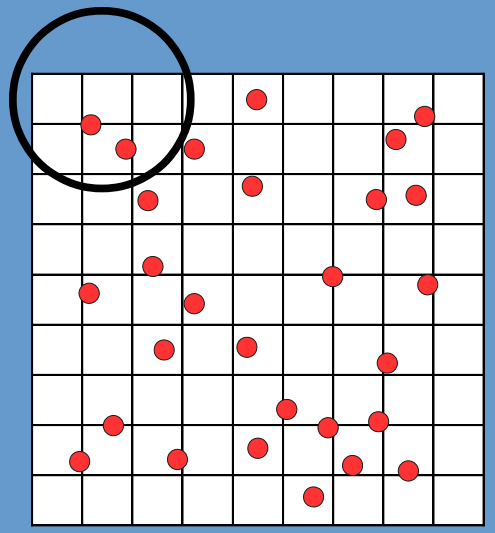
SaTScan

- Moving window method
  - Location of window (grid)



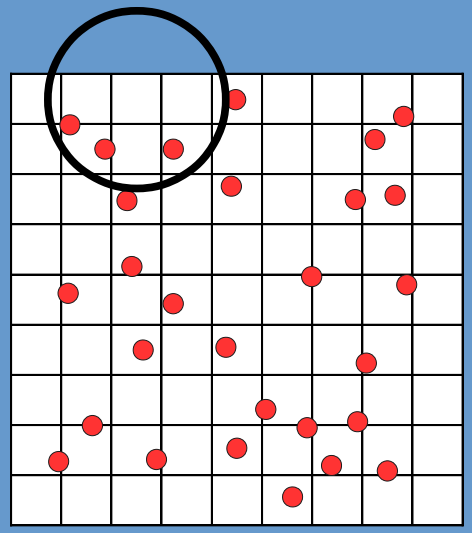
SaTScan

- Moving window method
  - Location of window (grid)



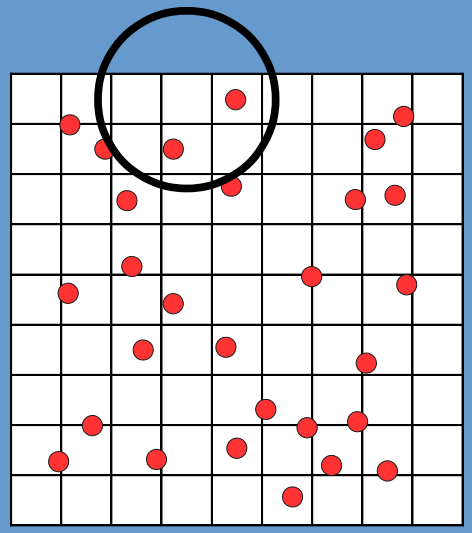
SaTScan

- Moving window method
  - Location of window (grid)



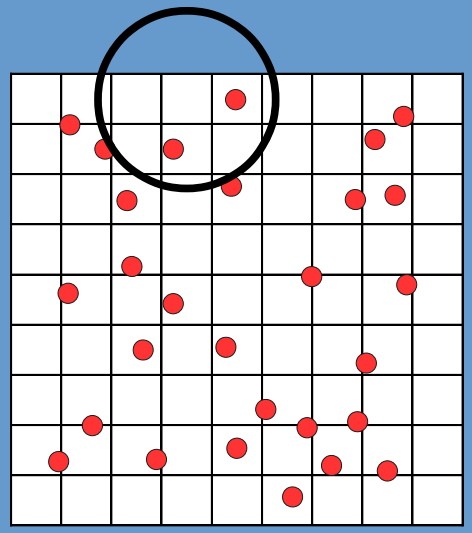
SaTScan

- Moving window method
  - Location of window (grid)



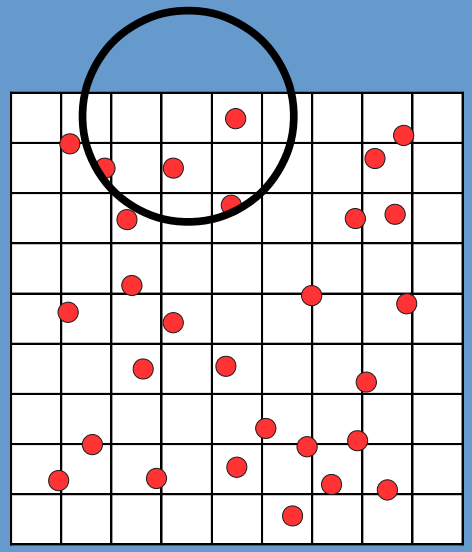
SaTScan

- Moving window method
  - Size of window (grid)



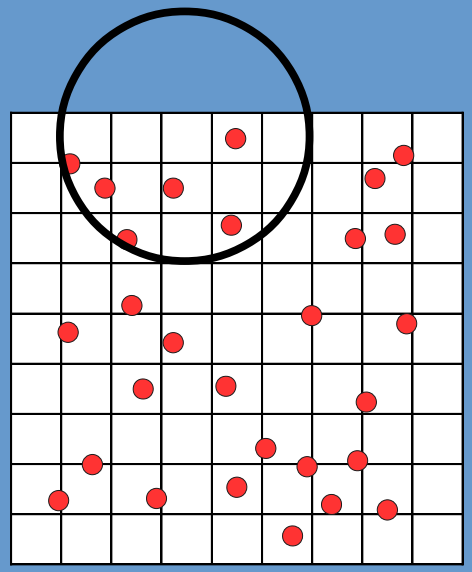
SaTScan

- Moving window method
  - Size of window (grid)



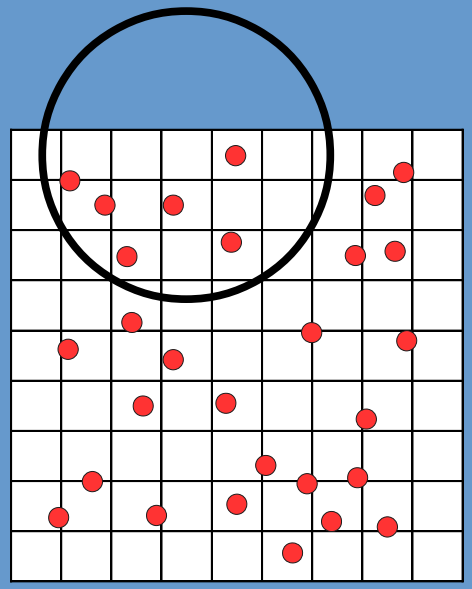
SaTScan

- Moving window method
  - Size of window (grid)



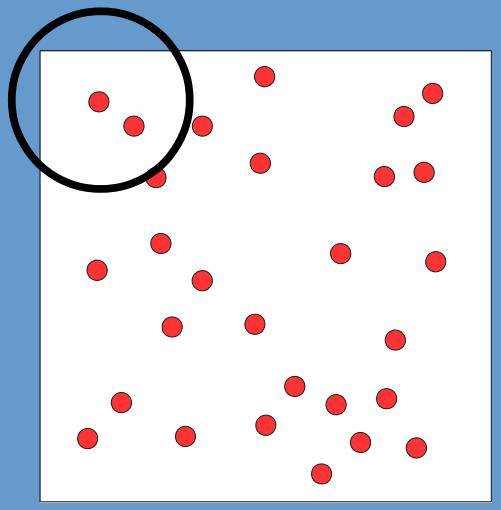
SaTScan

- Moving window method
  - Size of window (grid)



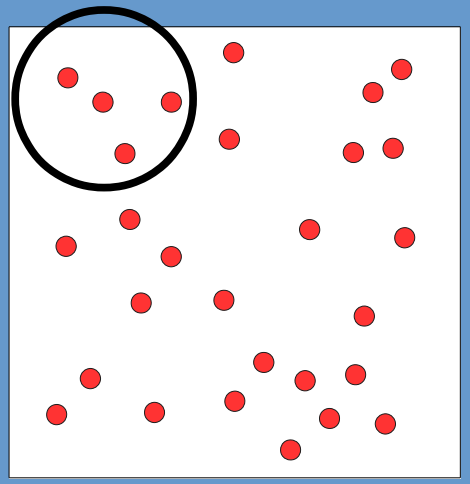
SaTScan

- Moving window method
  - Location of window (no grid)



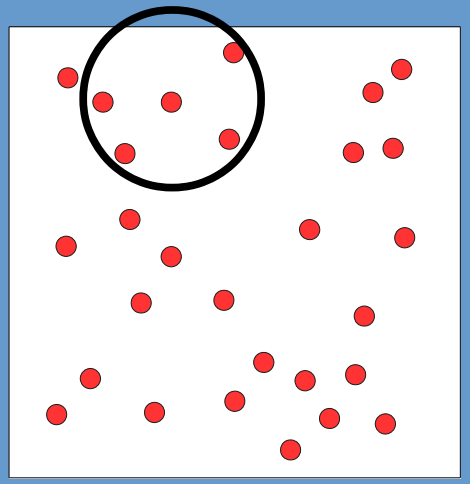
SaTScan

- Moving window method
  - Location of window (no grid)



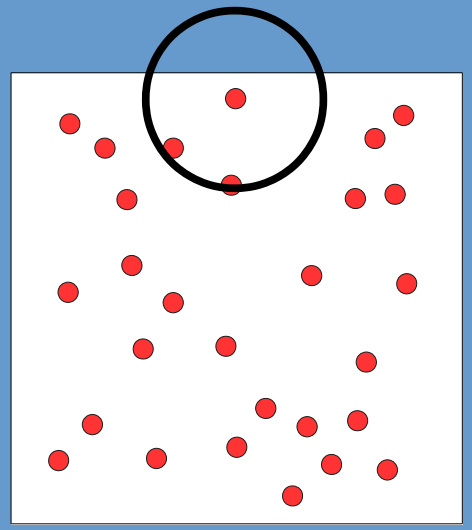
SaTScan

- Moving window method
  - Location of window (no grid)



SaTScan

- Moving window method
  - Location of window (no grid)



SaTScan

- Window size options
  - Allows us to set a threshold for cluster size
  - By total population size (%)
  - By physical size of window (distance)
    - Important to control for potentially large spatial clusters
      - Not really a "local cluster"?

SaTScan

- Window shape options
  - Circular (default)
  - Elliptical
    - Contains penalty option for deviation from circular
  - User-specified
    - Window defined by set of neighbors for each observation

SaTScan

- Spatiotemporal model (cases only)
  - Extend the spatial model with time
    - Rather than a circle in 2D (space), cluster is a cylinder in 3D (2D space plus time)
    - Basic idea is similar
      - Rather than simply moving a circle of various sizes across space, move a cylinder of various sizes across space and time
    - Can account for varying intensity through time

SaTScan

- Spatiotemporal model inputs (cases only)
  - Location coordinates
    - Each location has case, population count, and time
      - These are point representations
      - Thus, for areal (polygon) data, they need to be converted to points (centroids)
  - Prediction coordinates
    - Grid of locations (If no grid, locations are used)
    - Temporal window

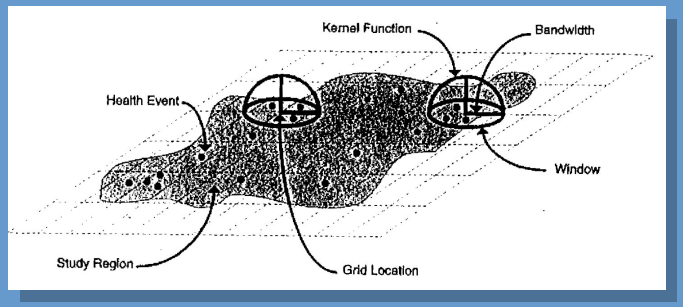
SaTScan

- Calculates a likelihood ratio for each potential cluster
  - Based on observed and expected disease cases
  - Adjusts for multiple tests (similar clusters)
    - Identifies most likely cluster
    - *e.g., consider the number of "high" clusters that would be found containing a single location with an extremely high count*

SaTScan

- Advantages
  - Often used/accepted
    - Mathematically sound, straightforward interpretation
  - Quite customizable for purpose
- Disadvantages
  - Spatial output options not great
  - Cluster shape restrictions

Density-based Clustering



- Point density surface
  - Requires (user input)
    - Regularly-spaced grid of locations
    - Kernel or density function
    - Kernel size
- Areal regions / extent of "cluster"
  - Requires (user input)
    - Areal region where point density > X

Density-based Clustering

- Advantages
  - Can use unmarked or marked point data
  - Density surfaces look nice
- Disadvantages
  - Subjectivity (kernel function, size)
  - Requires a user-defined threshold for identifying a “cluster”

Local Clusters

- Considerations
  - What is the question?
    - e.g., observations vs groups of obs
  - Exploratory or confirmatory
    - e.g., a priori beliefs about presence of a cluster?
  - Scale
    - What is the scale of a local cluster?... for your health outcome, in your region?

Local Clusters

- Considerations
  - What is the question?
    - e.g., observations vs groups of obs
  - Exploratory or confirmatory
    - e.g., a priori beliefs about presence of a cluster?
  - Scale
    - What is the scale of a local cluster?... for your health outcome, in your region?

LISA vs SatScan

- Considerations
  - Exploratory vs Confirmatory Analysis
    - LISA: Hypothesis Generation
    - SatScan: Hypothesis Testing
  - Individual-level vs group-level
    - LISA: How anomalous (outlier!) is each observation compared to the overall spatial patterns across the study area
    - SatScan: What is the 'group' of elevated or lower risk compared to expected?

Keywords

- Kulldorff
- SaTScan
- Moving window
  - Window size, shape