Spatially Significant Cluster Detection

Alan T Murray, Tony H Grubesic, Ran Wei

Published: Spatial Statistics, Vol 10, 2014, pp: 103-116

Introduction & Motivations

- The Problem:
  - "While spatial contiguity is widely considered an important condition of a cluster, most detection approaches employ a priori artificial structure, leading to disingenuous significance and unintended spatial biases that hinders meaningful discovery and interpretation."
- Critical sub-issues:
  - "The use of predefined geometric shapes can mask the actual morphology of [clusters]"
  - "Imposed structures can impede statistical inference, masking the underlying causes of clusters and their social, economic and ecological environment."
- This Paper:
  - Reviews the implications of assumed spatial structure
  - Develops a likelihood maximization approach without an assumed spatial window

Current Methodologies

- A range of different clustering methods have been proposed to reflect different concerns and interests, originating from researchers in mathematics & statistics, geography & GIS, criminology and epidemiology.
- Scanning methods are only contextually powerful/useful
- Two prominent methods used in cluster detection in Crime & Health contexts:
  - Spatial Scan Statistic
    - Believed to have high statistical power in identifying clusters, but imposes structure using (circular, elliptical) scan window
    - Imposed window structure may be misleading and introduce bias or lead to erroneous conclusions
  - Spatial Autocorrelation (Defined and not adapted for this paper)
    - Largely insensitive to cluster shape, but spatial structure is assumed via a neighborhood matrix w_{ij}

“Adaptations” of SaTScan

- Important (relevant) information:
  - Attempts of Identify Most Likely Cluster by identifying optimized LLR
  - Identifies a set of contiguous spatial units that has the highest probability of being a non-random spatial agglomeration
- Notation:
  - $R = \{1, 2, \ldots, N\}$ Total region under investigation
  - $\mathbf{Z} \subseteq R$ Window for scanning
  - $\Omega_{\mathbf{Z}} = \{i | i \in \mathbf{Z}\}$
  - n_i Population (or size) of region $i = 1, \ldots, N$
  - Y_i $\#$ of cases observed in region $i = 1, \ldots, N$
    note: we will assume $Y_i \sim Poisson\,(n_i \lambda_i)$

More Variables

- We will define the following variables for usage:
  - $Y_+ = \sum_i Y_i$ and $n_+ = \sum_i n_i$
  - $Y_{in} = \sum_{i \in \mathbf{Z}} Y_i$ and $n_{in} = \sum_{i \in \mathbf{Z}} n_i$
  - $\lambda_0 = \frac{Y_+}{n_+}$
  - $e_i = n_i \lambda_0$ (under H_0)
  - $e_{\mathbf{Z}} = \sum_{i \in \mathbf{Z}} n_i \lambda_0 = \lambda_0 \sum_{i \in \mathbf{Z}} n_i = \lambda_0 n_{in}$
- The Likelihood Ratio Statistic for the window $\mathbf{Z}$ is then:

$$LR(\mathbf{Z}) = \left(\frac{Y_{in}}{e_{\mathbf{Z}}} \right)^{Y_{in}} \left(\frac{Y_+ - Y_{in}}{Y_+ - e_{\mathbf{Z}}} \right)^{Y_+ - Y_{in}} I\left(Y_{in} > e_{\mathbf{Z}}\right)$$

Global and Local Autocorrelation (later considerations)

- Suppose that $|R| = N$, then we define $\bar{y} = \frac{Y_+}{N}$ and $z_i = y_i - \bar{y}$
- Let $[w]_{ij}$ be the spatial weights matrix indicating neighborhood structure
- Global Moran's I:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j z_i w_{ij} z_j}{\sum_i z_i^2}$$

- Local Moran's I (Local Indicator of Spatial Autocorrelation (LISA))

$$I_i = \frac{z_i}{\sum_i z_i^2} \sum_j w_{ij} z_j$$

Optimization

Theorem

Window z^* that maximizes the value of $(Y_{in} - e_{\mathbf{Z}})$ also maximizes $LLR(Z)$, for all windows with r-observed cases (i.e. $Y_{in} = r$)

$$LLR(\mathbf{Z}) = ln\left(\frac{Y_{in}}{e_{\mathbf{Z}}}\right)^{Y_{in}} + ln\left(\frac{Y_+ - Y_{in}}{Y_+ - e_{\mathbf{Z}}}\right)^{Y_+ - Y_{in}}$$

$$LLR(\mathbf{Z}) = Y_{in}\left[ln(Y_{in}) - ln(e_{\mathbf{Z}})\right] + (Y_+ - Y_{in})\left[ln(Y_+ - Y_{in}) - ln(Y_+ - e_{\mathbf{Z}}\right.$$

Theorem Proof

Proof.

We show that:

$$\frac{\partial}{\partial e_{\mathbf{Z}}}\left[LLR(\mathbf{Z})\right]=0 \iff (Y_{in}-e_{\mathbf{Z}})=0$$

$$\frac{\partial}{\partial e_{\mathbf{Z}}}\left[LLR(\mathbf{Z})\right]=\frac{\partial}{\partial e_{\mathbf{Z}}}\left[Y_{in}\left[ln(Y_{in})-ln(e_{\mathbf{Z}})\right]\right]$$

$$+\frac{\partial}{\partial e_{\mathbf{Z}}}\left[(Y_{+}-Y_{in})\left[ln(Y_{+}-Y_{in})-ln(Y_{+}-e_{\mathbf{Z}})\right]\right]$$

$$=\frac{\partial}{\partial e_{\mathbf{Z}}}\left[-Y_{in}ln(e_{\mathbf{Z}})\right]+\frac{\partial}{\partial e_{\mathbf{Z}}}\left[-(Y_{+}-Y_{in})\,ln(Y_{+}-e_{\mathbf{Z}})\right]$$

$$=\frac{-Y_{in}}{e_{\mathbf{Z}}}+\frac{Y_{+}-Y_{in}}{Y_{+}-e_{\mathbf{Z}}}=\frac{Y_{+}(e_{\mathbf{Z}}-Y_{in})}{e_{\mathbf{Z}}(Y_{+}-e_{\mathbf{Z}})}=0 \iff (e_{\mathbf{Z}}-Y_{in})=0$$

□

Linear Optimization: Non-Contiguous

- We can exploit the results of the previous theorem to find an optimal window without assuming any spatial structure using a linear optimization model. We define the following:
  - $i =$ index of spatial units for $i \in \{1, \dots, N\}$
  - $Y_{(i)}$ for $i = 1, \dots, N$ the ordered response values of Y_i
  - T_j for $j = 1, \dots, m$ the unique values of $Y_{(i)}$ where $m \leq N$
  - $r_1 = \{Y_+\}$
    $r_2 = \{(Y_+ - T_{j_2}) | \ j_2 \in \{1, \dots, m\}\}$
    $r_3 = \{(Y_+ - (T_{j_2} + T_{j_3})) | \ j_2 \neq j_3 \in \{1, \dots, m\}\}$
    $\vdots$
    $r_m = \left\{ Y_+ - \sum_{j_p \neq \{j_k\}_{k=3}^{m-1}} T_{j_p} \right\}$ $\qquad r = \bigcup_{k=1}^{m} r_k$
  - $X_i = \begin{cases} 1 & \text{if} \quad i \in \mathbf{Z}^* \\ 0 & \text{if} \quad i \notin \mathbf{Z}^* \end{cases}$

Linear Optimization (Continued)

- We can then find the window $\mathbf{Z}^*$ that maximizes the likelihood function over r by solving the optimization model:
  - Maximize: $\sum_i (y_i - e_i) X_i$
  - Subject to: $\sum_i Y_i X_i = r_k$ for $k = 1, \ldots, m$
  - $X_i \in \{0, 1\} \quad \forall i$

- Note: There is no guarantee that the selected units, or the cluster, will be spatially contiguous.

Ensuring Spatial Contiguity

- More Notation!
  - N_i Set of Spatial Units adjacent to i
  - M A large value that is set to the total number of spatial units
  - F_{ij} Amount of flow between units $i \to j$
  - $V_i = \begin{cases} 1 & \text{if} \quad i \textit{ is a sink} \\ 0 & \text{if} \quad i \textit{ otherwise} \end{cases}$
- Then the linear optimization problem becomes:
  - Maximize: $\sum_i (y_i - e_i) X_i$
  - Subject to: $\sum_i Y_i X_i = r_k$ for $k = 1, \ldots, m$
  - $\sum_{j \in N_i} F_{ij} - \sum_{j \in N_i} F_{ji} \geq X_i - MV_i \ \ \forall i$ and $\sum_i V_i = 1$
  - $\sum_{j \in N_i} Y_{ij} \leq (M-1) X_i \ \ \forall i, \qquad V_i \leq X_i \ \ \forall i$
  - $X_i \in \{0, 1\} \quad \forall i \qquad$ and $V_i \in \{0, 1\} \quad \forall i$

Cincinnati Assault Analysis

- Background Information
  - Observations from 457 Census blocks in several different neighborhoods in Cincinnati
  - Data collected for the first six months of 2008
  - Total number of assault cases in area was 462
  - Population at time of observation was approximately 38,700 over 148.7 km^2
  - Total assault rate for the region is approximately 1.6 times the rate of the overall Cincinnati area, with relative risk ranging from 0-45.2

Study Area

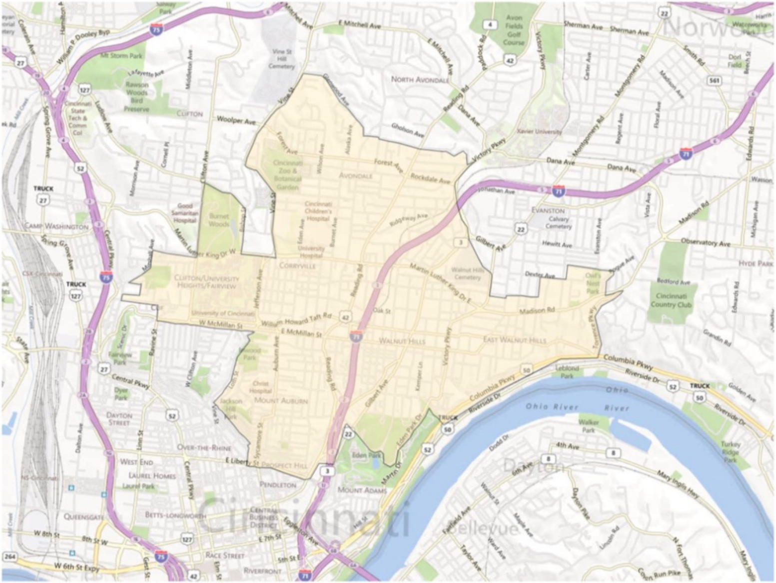

Fig. 1. Study area (neighborhoods—Clifton, Walnut Hills, Evanston and Avondale).

SaTScan Results

Table 1
Spatial scan statistic findings (derived by SaTScan).

Max window size	Actual cases (c_Z)	Expected cases (μ_Z)	Total selected units	LLR
500 feet	17	1.15	3	30.24
1000 feet	24	2.78	6	31.09
3000 feet	119	53.67	93	35.49
5000 feet	289	174.69	254	62.07

Fig. 2a. Spatial configuration of spatial scan statistic with maximal window size 1000 ft.

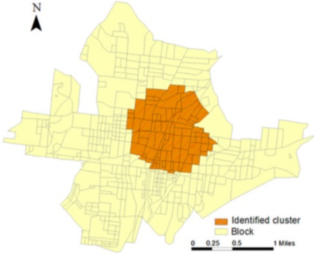

Fig. 2b. Spatial configuration of spatial scan statistic with maximal window size 3000 ft.

Max LLR Results

Table 2
Max-LLR model findings.

Actual cases (c_Z)	Expected cases (μ_Z)	Total selected units	LLR
17	0.71	5	38.07
24	1.16	6	50.48
119	16.02	39	149.75
289	85.11	93	227.79

Fig. 3a. Spatial configuration of Max-LLR with 24 observed cases.

Fig. 3b. Spatial configuration of Max-LLR with 119 observed cases.

Contiguous Max LLR Results

Table 3
Contiguous-Max-LLR model findings.

Actual cases (c_Z)	Expected cases (μ_Z)	Total selected units	LLR
17	1.15	4	30.24
24	2.37	21	34.53
119	20.13	61	125.65
289	91.14	120	210.48

Fig. 4a. Spatial configuration of Contiguous-Max-LLR with 24 observed cases.

Fig. 4b. Spatial configuration of Contiguous-Max-LLR with 119 observed cases.