Optimal Parameters for Space-Time Cluster Detection of Infectious Disease

Evan Caten
Masters Candidate
Salem State College
May 4, 2009
Presentation Outline

- Overview of masters thesis

  - Introduction
  - Objectives
  - Approach & Methods
  - Results
  - Challenges & Limitations
  - Recommendations for Future Research
  - Questions
The reality of healthcare… is that geography is destiny.”

19th century colonialism, trade & exotic disease

Disease & environment
- Exploring relationships

Mapping for disease or maps as report adjuncts

Public Health & Spatial Epidemiology

Public Health & Spatial Epidemiology

- Advent of computer aided cartography leads to automated process & later – GIS.

- “Spatial statistics” and recognition of spatial properties.

- Early “clustering” algorithms (Moran’s I) apply rigorous statistical methods.
Spatial Analysis & “Clustering”

- “a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance”
  
  Knox 1989

- “any area within the study region of significant elevated risk”
  
  Lawson 2001
Simplified Epidemiology Cluster Definition

Epidemiology Model:
- Close Contact with Case
- Within Infectious Period
  * many additional factors
Automated “scanning window” algorithms find areas of high clustering

Geographical Analysis Machine (GAM)

The more recent SaTScan software enhances this general model
Health Data & Aggregation

Confidentiality & Privacy

- Suppression often required (HIPAA, FIPAA regulations)
- De-identification by aggregation using common systems
- Management of large disease volume

Census geography provides convenient aggregation zones for de-identification & data management
Health Data & Aggregation

- Issues with Aggregation
  - Ecological fallacy & heterogeneity
  - MAUP
    - Ability to misrepresent with choice of zones
    - Choice of scale affects pattern
  - Edge effects: the non-uniformity of space

Different aggregations of Snow's data present different pictures (Monmonier 1996)
Objectives

- Find Optimal Parameters
  - To find optimal spatial parameters for use in automated cluster detection software.

  “Which aggregate zonal system performs best spatially and temporally for infectious disease cluster detection?”

- Recommendations for Public Health
  - Cataloging the effects of aggregation & scale on cluster detection.
  - Seeks to add to the knowledge base & provide guidance in the use of automated detection systems to support cluster detection.
Approach & Methods

- Extract Previously Identified Clusters
- Define significant signals with software
- Match clusters & signals
- Performance evaluation
Preliminary Steps

- **Dataset Requirements**
  - Contains previously identified clusters with known spatial & temporal dimensions.
  - Three consecutive years representing all of Massachusetts
    - To minimize any seasonal and regional bias
  - Complete set of spatial identifiers (block-group, tract, zip, city)
    - To ensure equal participation in all aggregate analyses
  - Full and valid calendar date of event
    - To remove any temporal rounding error
  - Analysis ready dataset with > 1000 records.
    - To minimize small-number problems

- **Pertussis Dataset 2005-2007**
  - Limited dataset agreement
Dataset Preparation & Cleaning

- **Range Checks**
  - Date Values
  - Zone Values
  - Zonal Cross-Reference
  - Non-Areal Zips Mapped

- **Incomplete Records**
  - 13 records with incomplete zones

---

**Limited dataset description**

<table>
<thead>
<tr>
<th>Observations</th>
<th>Complete* Observations</th>
<th>Unique Clusters</th>
<th>Temporal Range</th>
</tr>
</thead>
</table>

**Variables**

Disease, Case Status, Event Date, City Name, Zip-Code, Census Tract, Census Block-Group, Cluster ID

*13 records missing at least one value
Known Cluster Extraction

- **Group by Cluster ID**
  - 550 records
  - 94 clusters

- **Temporal Dimensions**
  - First & Last Date = Start & End
  - Difference = Duration

- **Spatial Dimensions (for BGP, TRC, ZIP, TWN)**
  - Weighted center (counts by zone centroid)
  - Area = Sum of component zones

- **Population & Event Count**

<table>
<thead>
<tr>
<th>Known Cluster Duration</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>46.71</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>31.05</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>963.95</td>
</tr>
<tr>
<td>Range</td>
<td>143</td>
</tr>
<tr>
<td>Minimum</td>
<td>1</td>
</tr>
<tr>
<td>Maximum</td>
<td>144</td>
</tr>
<tr>
<td>Count</td>
<td>94</td>
</tr>
</tbody>
</table>

Descriptive statistics on temporal duration of known clusters
Previously identified clusters by zone type

A. Block-Group

Clusters: 94
Avg. Durn.: 46.71 (days)
Avg. Zones: 4.20 (n)
Avg. Cases: 5.85 (n)
Avg. Area: 19.58 (SqKm)
Avg. Pop: 6,197.50 (n)

B. Tract

Clusters: 94
Avg. Durn.: 46.71 (days)
Avg. Zones: 2.98 (n)
Avg. Cases: 5.85 (n)
Avg. Area: 51.55 (SqKm)
Avg. Pop: 15,798.00 (n)

C. ZipCode

Clusters: 94
Avg. Durn.: 46.71 (days)
Avg. Zones: 1.93 (n)
Avg. Cases: 5.85 (n)
Avg. Area: 71.80 (SqKm)
Avg. Pop: 41,031.35 (n)

D. Town

Clusters: 94
Avg. Durn.: 46.71 (days)
Avg. Zones: 1.71 (n)
Avg. Cases: 5.85 (n)
Avg. Area: 96.88 (SqKm)
Avg. Pop: 125,318.90 (n)
## Known cluster average indicators by zone type

<table>
<thead>
<tr>
<th>Zone Type</th>
<th>Total Clusters</th>
<th>Average Duration (days)</th>
<th>Average Zone Count</th>
<th>Average Case Count</th>
<th>Average Area (SqKm)</th>
<th>Average Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block-Group</td>
<td>94</td>
<td>46.71</td>
<td>4.20</td>
<td>5.85</td>
<td>19.58</td>
<td>6,197.50</td>
</tr>
<tr>
<td>Tract</td>
<td>94</td>
<td>46.71</td>
<td>2.98</td>
<td>5.85</td>
<td>51.55</td>
<td>15,798.00</td>
</tr>
<tr>
<td>Zip-Code</td>
<td>94</td>
<td>46.71</td>
<td>1.93</td>
<td>5.85</td>
<td>71.80</td>
<td>41,031.35</td>
</tr>
<tr>
<td>City/Town</td>
<td>94</td>
<td>46.71</td>
<td>1.71</td>
<td>5.85</td>
<td>96.88</td>
<td>125,318.90</td>
</tr>
</tbody>
</table>
Signal Definition

- **Space-time scanning window (SaTScan algorithm)**
  - Irregular lattice from zone centroids
  - Pop & events aggregated by zone
  - Likelihood-ratio statistic in-zone vs outside

Scan windows center on point and expand in 3 dimensions creating # of windows to test then moves to next centroid and repeats
Signal Definition

- Iterations by BGP, TRC, ZIP, TWN

- Individual maximum settings derived from known clusters
  - Eliminates oversized clusters

- Time aggregation by pertussis Infectious period (21 days)
  - Improves performance

- Software output contains indicators
  - Component zones, population, total events, start/end dates.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Cluster Length (days)</td>
<td>96</td>
</tr>
<tr>
<td>Max Cluster Radius-Census block-Group (km)</td>
<td>7</td>
</tr>
<tr>
<td>Max Cluster Radius-Census Tract (km)</td>
<td>11</td>
</tr>
<tr>
<td>Max Cluster Radius-Zip-Code (km)</td>
<td>12</td>
</tr>
<tr>
<td>Max Cluster Radius-City/Town (km)</td>
<td>14</td>
</tr>
</tbody>
</table>

SaTScan software cluster detection maximum limits derived from known clusters
Significant signals ($p<.05$) by zone type

A. Block-Group
- Clusters: 19
- Avg. Durn.: 65.47 (days)
- Avg. Zones: 24.79 (n)
- Avg. Cases: 16.74 (n)
- Avg. Area: 53.97 (SqKm)
- Avg. Pop: 32,447.79 (n)

B. Tract
- Clusters: 21
- Avg. Durn.: 67.14 (days)
- Avg. Zones: 6.71 (n)
- Avg. Cases: 17.00 (n)
- Avg. Area: 94.72 (SqKm)
- Avg. Pop: 35,761 (n)

C. ZipCode
- Signals: 23
- Avg. Durn.: 68.39 (days)
- Avg. Zones: 3.52 (n)
- Avg. Cases: 17.17 (n)
- Avg. Area: 156.55 (SqKm)
- Avg. Pop: 40,752.26 (n)

D. Town
- Signals: 19
- Avg. Durn.: 64.21 (days)
- Avg. Zones: 3.84 (n)
- Avg. Cases: 18.84 (n)
- Avg. Area: 216.81 (SqKm)
- Avg. Pop: 52,962.47 (n)
### Significant signal average indicators by zone type (p<=.05)

<table>
<thead>
<tr>
<th>Zone Type</th>
<th>Total Signals</th>
<th>Average Duration (days)</th>
<th>Average Zone Count</th>
<th>Average Case Count</th>
<th>Average Area (SqKm)</th>
<th>Average Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block-Group</td>
<td>19</td>
<td>65.47</td>
<td>24.79</td>
<td>16.74</td>
<td>53.97</td>
<td>32,447.79</td>
</tr>
<tr>
<td>Tract</td>
<td>21</td>
<td>67.14</td>
<td>6.71</td>
<td>17.00</td>
<td>94.72</td>
<td>35,761.10</td>
</tr>
<tr>
<td>Zip-Code</td>
<td>23</td>
<td>68.39</td>
<td>3.52</td>
<td>17.17</td>
<td>156.55</td>
<td>40,752.26</td>
</tr>
<tr>
<td>City/Town</td>
<td>19</td>
<td>64.21</td>
<td>3.84</td>
<td>18.84</td>
<td>216.81</td>
<td>52,962.47</td>
</tr>
</tbody>
</table>
Matching Clusters & Signals

- Space-time merge using temporal window and distance between centers.

\[ \exists \text{match}_i \left( [\text{ClusStrt}_c, \text{ClusEnd}_c] \cap [\text{SignStrt}_s, \text{SignEnd}_s] \neq \emptyset \right) \]

- Location intersection buffer varies by zone type
  - Derived from average radius (+/- 2SD) of zone type

- Temporal intersection buffer = 1 infectious period
  - Pertussis team definition: +/- 21 days

Spatial calculations in ArcGIS and subsequent temporal merge in MS Access
Matched clusters & signals by zone type

A. Block-Group

Matches: 8
Diff. Avg. Dist: 1.70 km
Diff. Avg. Start: 9.00 days

B. Tract

Matches: 11
Diff. Avg. Dist: 3.74 km
Diff. Avg. Start: 5.55 days

C. ZipCode

Matches: 7
Diff. Avg. Dist: 4.34 km
Diff. Avg. Start: 8.86 days

D. Town

Matches: 10
Diff. Avg. Dist: 3.62 km
Diff. Avg. Start: 7.46 days
Performance Evaluation

- **Sensitivity (ability to accurately detect clusters)**
  - Percent clusters detected \(\left(\frac{\text{signals}_n}{\text{clusters}_n}\right) \times \%\)

- **Specificity (temporal):**
  - \(\Delta\) in start dates \(\text{abs}([\text{ClusStrt}_i - \text{SignStrt}_i]) <\text{days}\)
  - \(\Delta\) in duration \(\text{abs}([\text{SignDurn}_i - \text{ClusDurn}_i]) <\text{days}\)

- **Specificity (spatial):**
  - \(\Delta\) between centers: \(d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}\) <\text{km}\)
  - \(\Delta\) in area: \(\text{abs}([\text{ClusArea}_i - \text{SignArea}_i]) <\text{km}^2\)
  - \(\Delta\) in component zones: \(\text{abs}([\text{ClusZones}_i - \text{SignZones}_i]) <\text{n}\)
  - \(\Delta\) in case counts: \(\text{abs}([\text{ClusCases}_i - \text{SignCases}_i]) <\text{n}\)
  - \(\Delta\) in population-at-risk: \(\text{abs}([\text{ClusPop}_i - \text{SignPop}_i]) <\text{n}\)
### Matched signals and clusters average indicator differences by zone type

<table>
<thead>
<tr>
<th>Zone Type</th>
<th>Total Matches</th>
<th>Sensitivity (% Match)</th>
<th>Average Start Date Difference (days)</th>
<th>Average Duration Difference (days)</th>
<th>Average Distance Difference (Km)</th>
<th>Average Area Difference (Sq Km)</th>
<th>Average Zone Count Difference</th>
<th>Average Case Count Difference</th>
<th>Average Population Difference (x1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block-Group</td>
<td>8</td>
<td>8.51</td>
<td>9.00</td>
<td>10.38</td>
<td>1.70</td>
<td>20.71</td>
<td>7.88</td>
<td>3.13</td>
<td>8.85</td>
</tr>
<tr>
<td>Tract</td>
<td>11</td>
<td>11.70</td>
<td>5.55</td>
<td>10.00</td>
<td>3.74</td>
<td>76.46</td>
<td>3.45</td>
<td>7.55</td>
<td>18.33</td>
</tr>
<tr>
<td>Zip-Code</td>
<td>10</td>
<td>10.64</td>
<td>8.86</td>
<td>9.90</td>
<td>4.34</td>
<td>95.99</td>
<td>2.70</td>
<td>10.80</td>
<td>25.25</td>
</tr>
<tr>
<td>City/Town</td>
<td>7</td>
<td>7.45</td>
<td>7.20</td>
<td>12.00</td>
<td>3.62</td>
<td>101.01</td>
<td>2.14</td>
<td>13.14</td>
<td>33.54</td>
</tr>
</tbody>
</table>

Averaging indicator differences by zone type yields a set of values used to evaluate performance across zone types.
Results

- **Sensitivity**
  - Low for all zone types (range 7.45% - 11.40%)
  - Tract performed marginally better than others
  - Percent of common matches was high (54% in >=3 zones)

- **Specificity**
  - Spatial & temporal specificity mixed across zone types
  - Overall tract performed best and city/town worst (ranking)
  - Predictable trending due to average zone size
Average indicator differences by aggregate zone type
Challenges & Limitations

- **Results don’t support any conclusions**
  - Very low match rate suggests that matching clusters and signals is a limiting factor

- **Modeling & definition of clusters**
  - Pertussis team uses a network model defining clusters based on “close-contact” and infectious period.
  - SaTScan uses a purely mathematical model based on space and time.

- **Data collection & reporting of infectious disease**
  - This analysis is based on home address.
  - No accounting for cluster type or source(s) of diffusion.
Recommendations for Future Research

- Cluster definition – aligning models
  - Cluster type classification
  - Investigate current matches for explaining factors (PCA)

- Data collection strategies
  - Additional data collection & development
  - Cluster type classification system (e.g. single source, linear, unknown?)

- Other diseases, longer timeframes.

*Automated systems may help support and enhance traditional detection models.*
Acknowledgements

Thesis Committee:

- Dr. William Hamilton, PhD (Advisor) - Salem State College
- Dr. John Hayes, PhD - Salem State College
- Mrs. Dawn Heisey-Grove, MPH – MA Dept of Public Health

Thanks to:

- Dr. Marcos Luna, SSC
- Ms. Tracy Laporte, MDPH
- Ms. Noelle Cocoros, MDPH
- Ms. Gillian Haney, MDPH
Questions?

Evan Caten
Salem State College
Department of Geography
ebcaten@hotmail.com