Optimal Parameters for Space-Time Cluster Detection of Infectious Disease

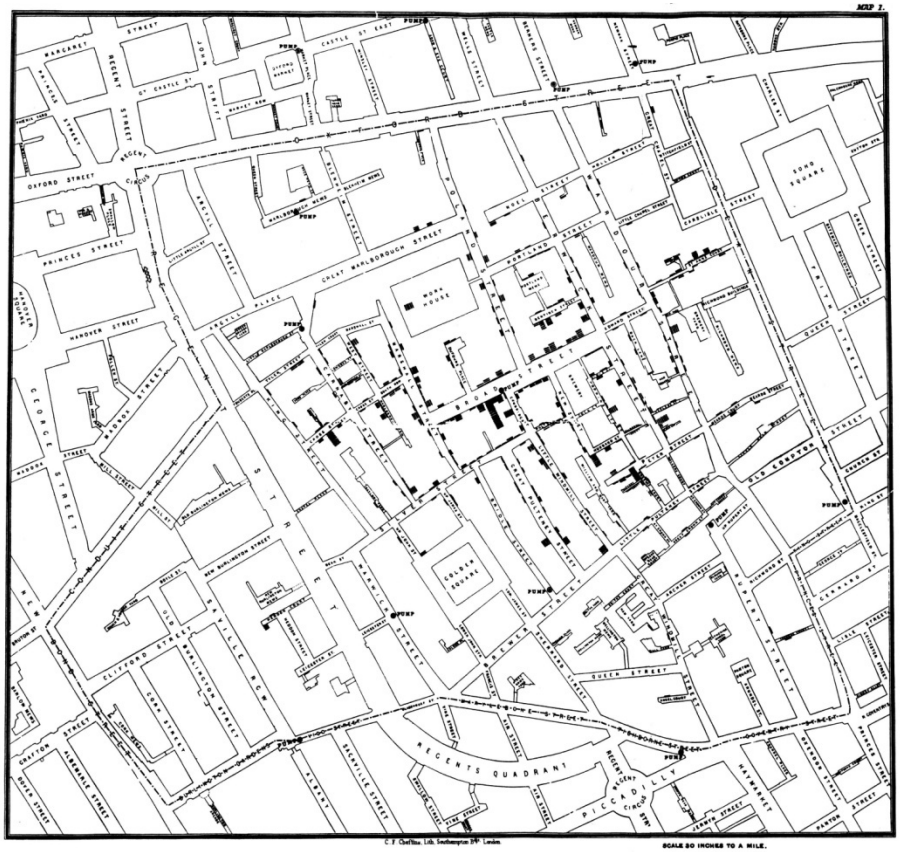
Evan Caten
Masters Candidate
Salem State College
May 4, 2009

Presentation Outline

- Overview of masters thesis
  - Introduction
  - Objectives
  - Approach & Methods
  - Results
  - Challenges & Limitations
  - Recommendations for Future Research
  - Questions

Public Health & Spatial Epidemiology

- 19th century colonialism, trade & exotic disease
- Disease & environment
  - Exploring relationships
- Mapping *for* disease or maps as report adjuncts



The reality of healthcare… is that geography is destiny."

Dartmouth Atlas of Health Care 1998.

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”

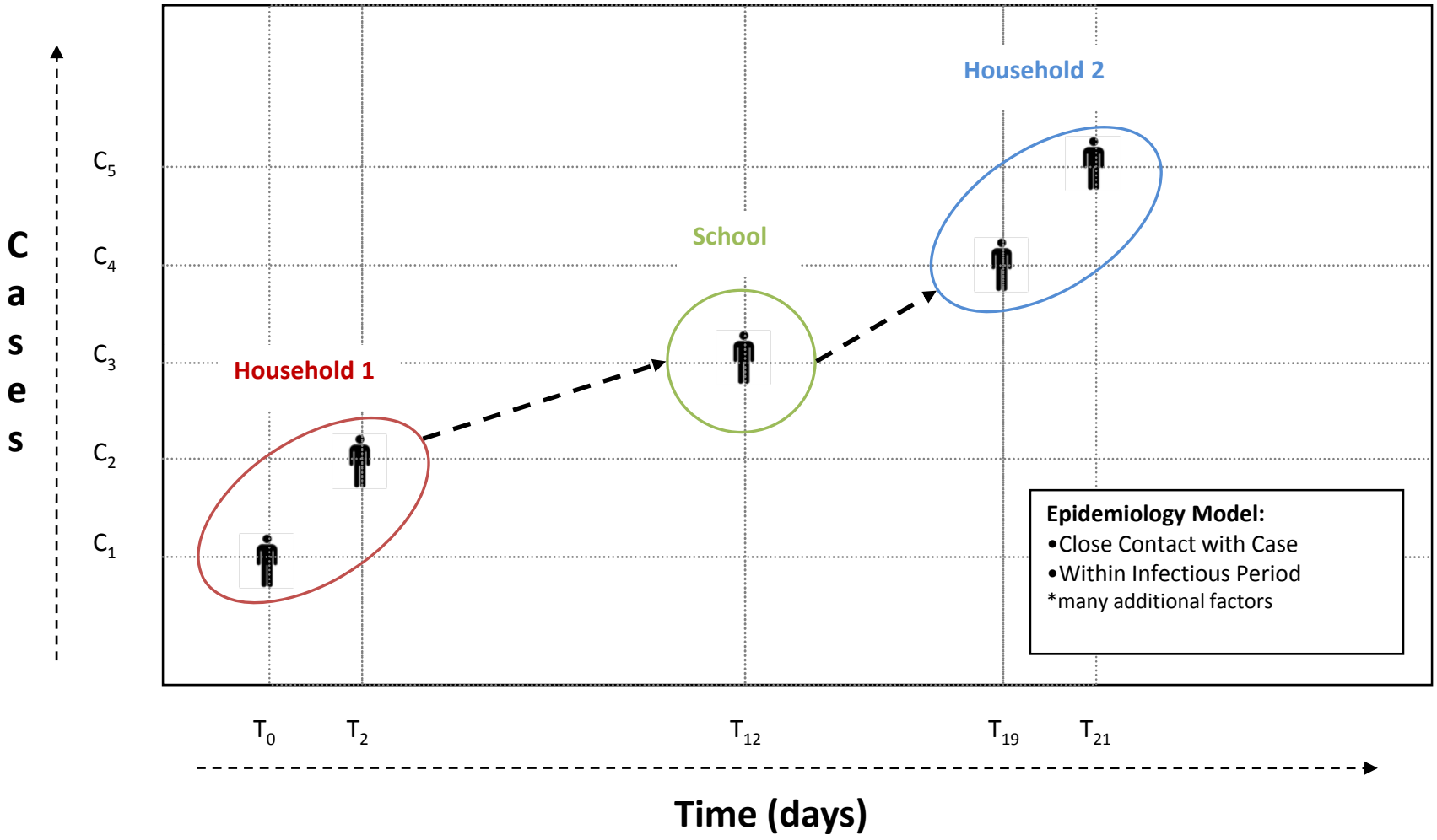
- “a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance”

  Knox 1989

  - “any area within the study region of *significant* elevated risk”

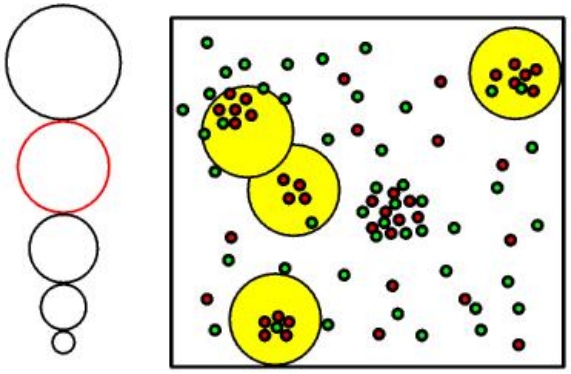
    Lawson 2001

Simplified Epidemiology Cluster Definition

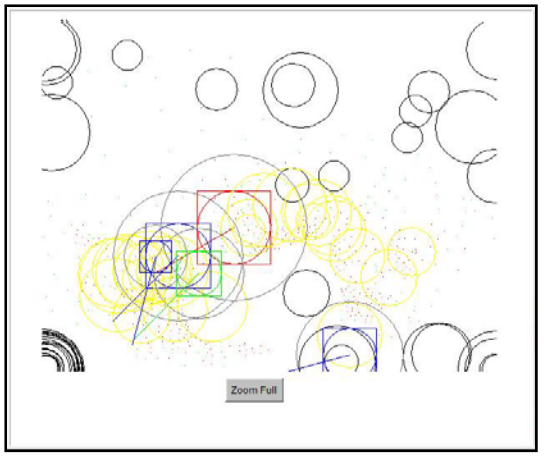


Mathematical Cluster Definition

- Automated “scanning window” algorithms find areas of high clustering
- Geographical Analysis Machine (GAM)
- The more recent SaTScan software enhances this general model



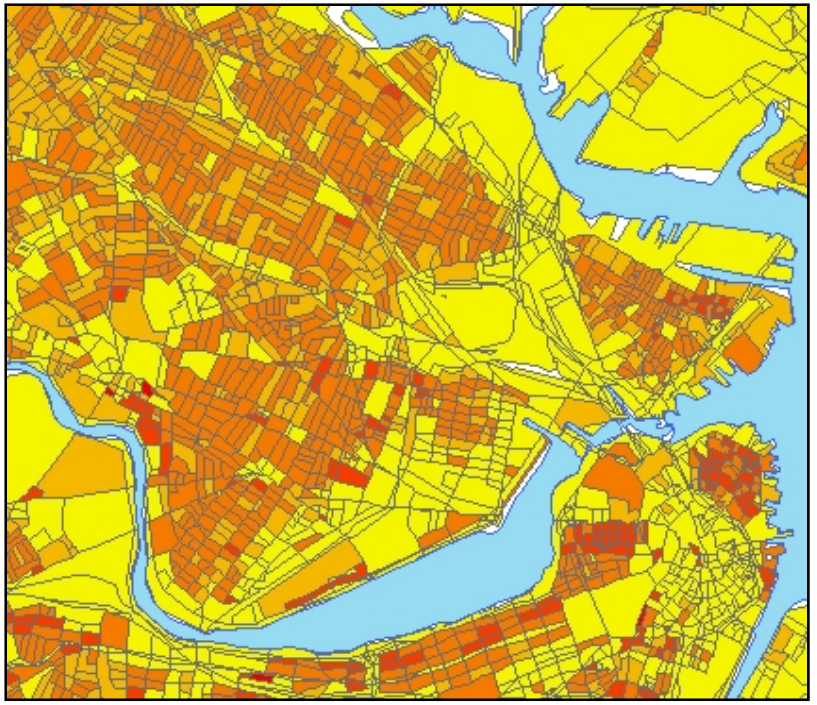
Fixed circular windows evaluate multiple variables for probability at each point



Fixed circular windows evaluate multiple variables for probability at each point

Health Data & Aggregation

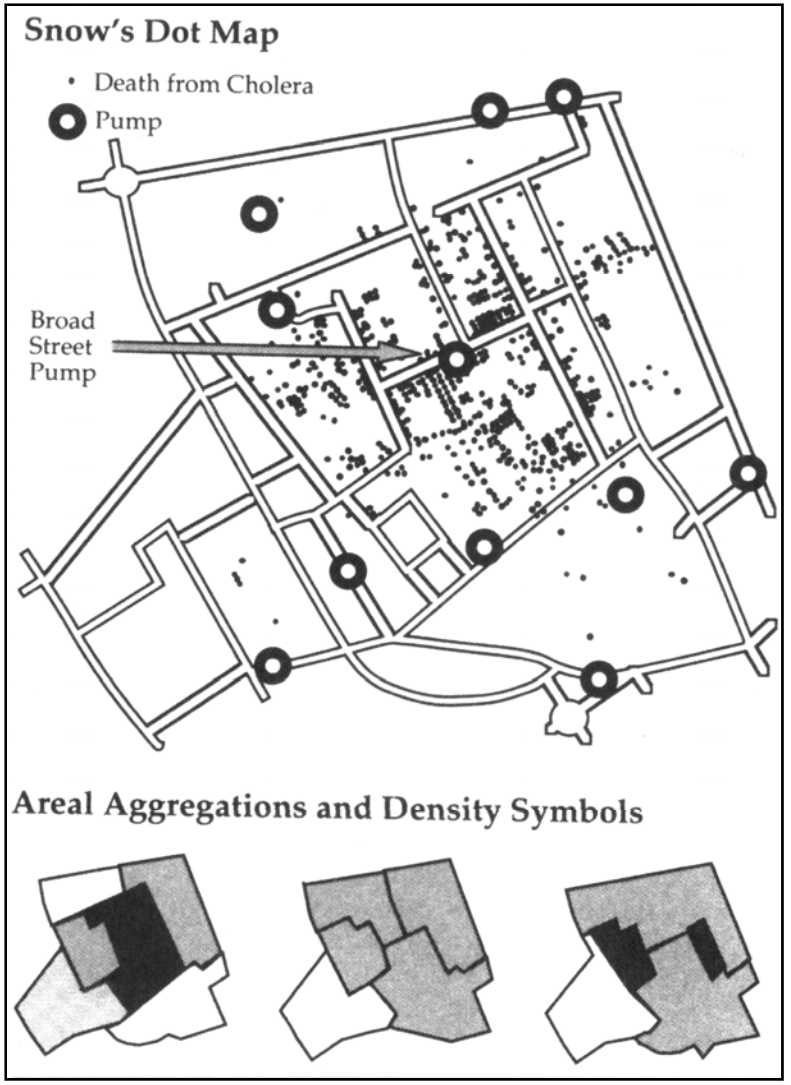
- Confidentiality & Privacy
  - Suppression often required (HIPAA, FIPAA regulations)
  - De-identification by aggregation using common systems
  - Management of large disease volume



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

Observations	Complete* Observations	Unique Clusters	Temporal Range
3,617	3,604	94	1/1/2005 - 12/31/2007

Variables

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

**13 records missing at least one value*

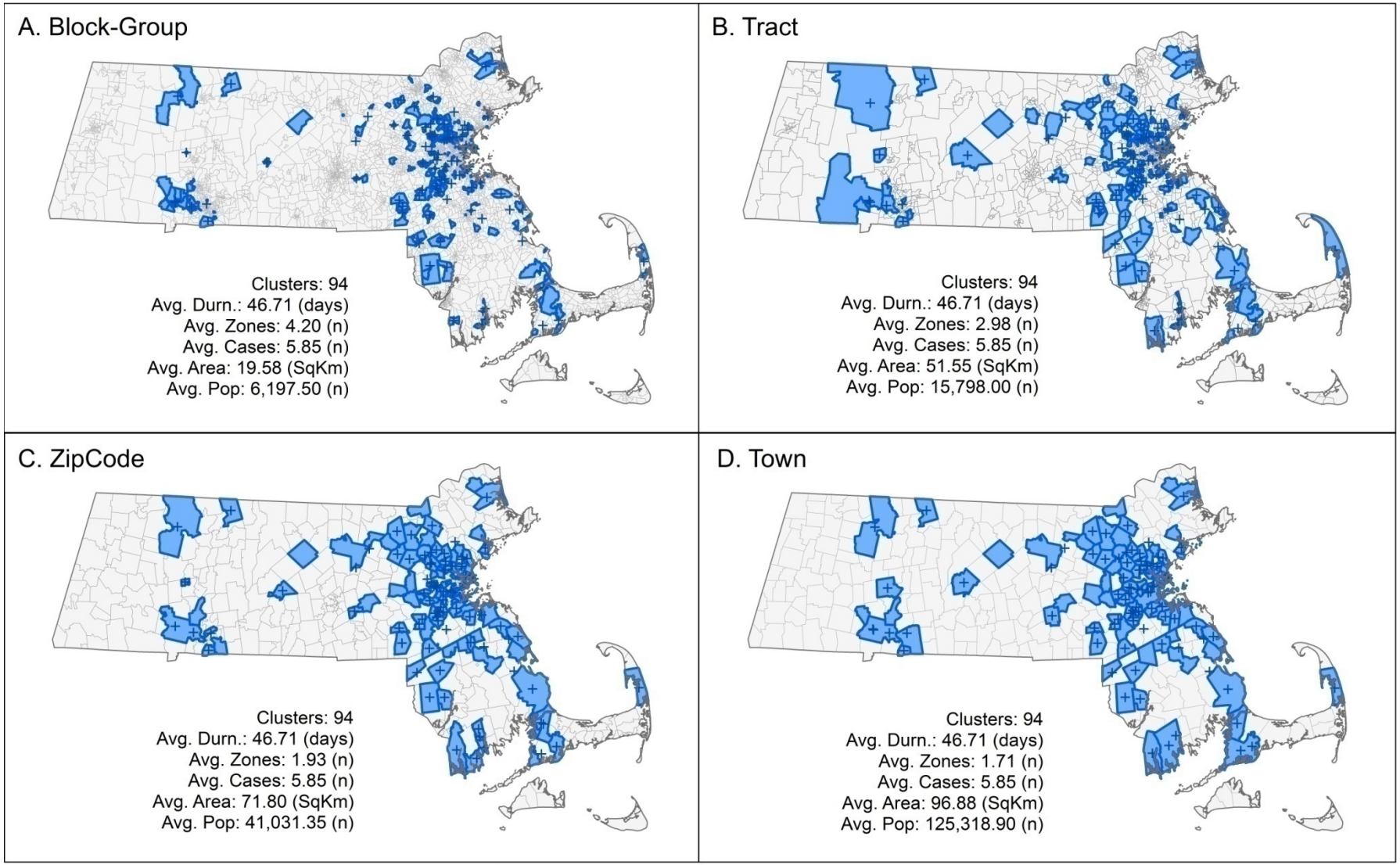
Known Cluster Extraction

- 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

Known Cluster Duration	
Mean	46.71
Standard Deviation	31.05
Sample Variance	963.95
Range	143
Minimum	1
Maximum	144
Count	94

Descriptive statistics on temporal duration of known clusters

Previously identified clusters by zone type

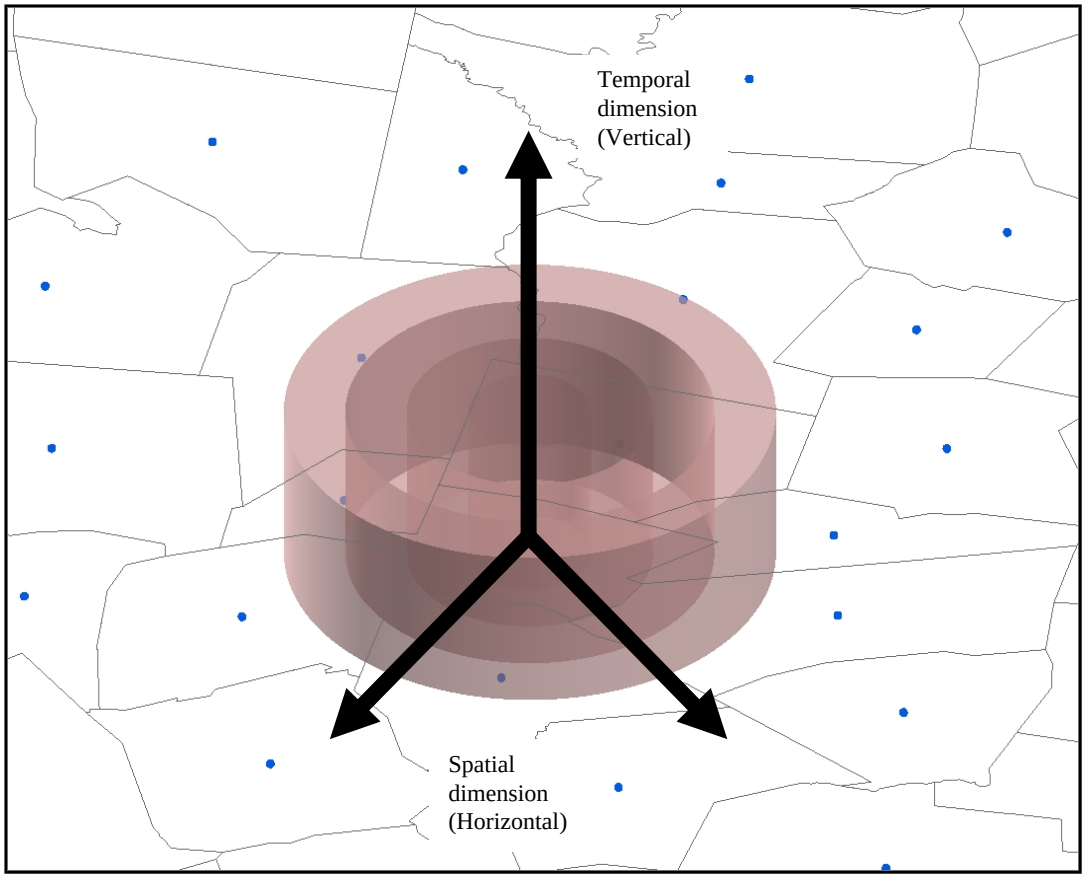


Known cluster average indicators by zone type

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

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

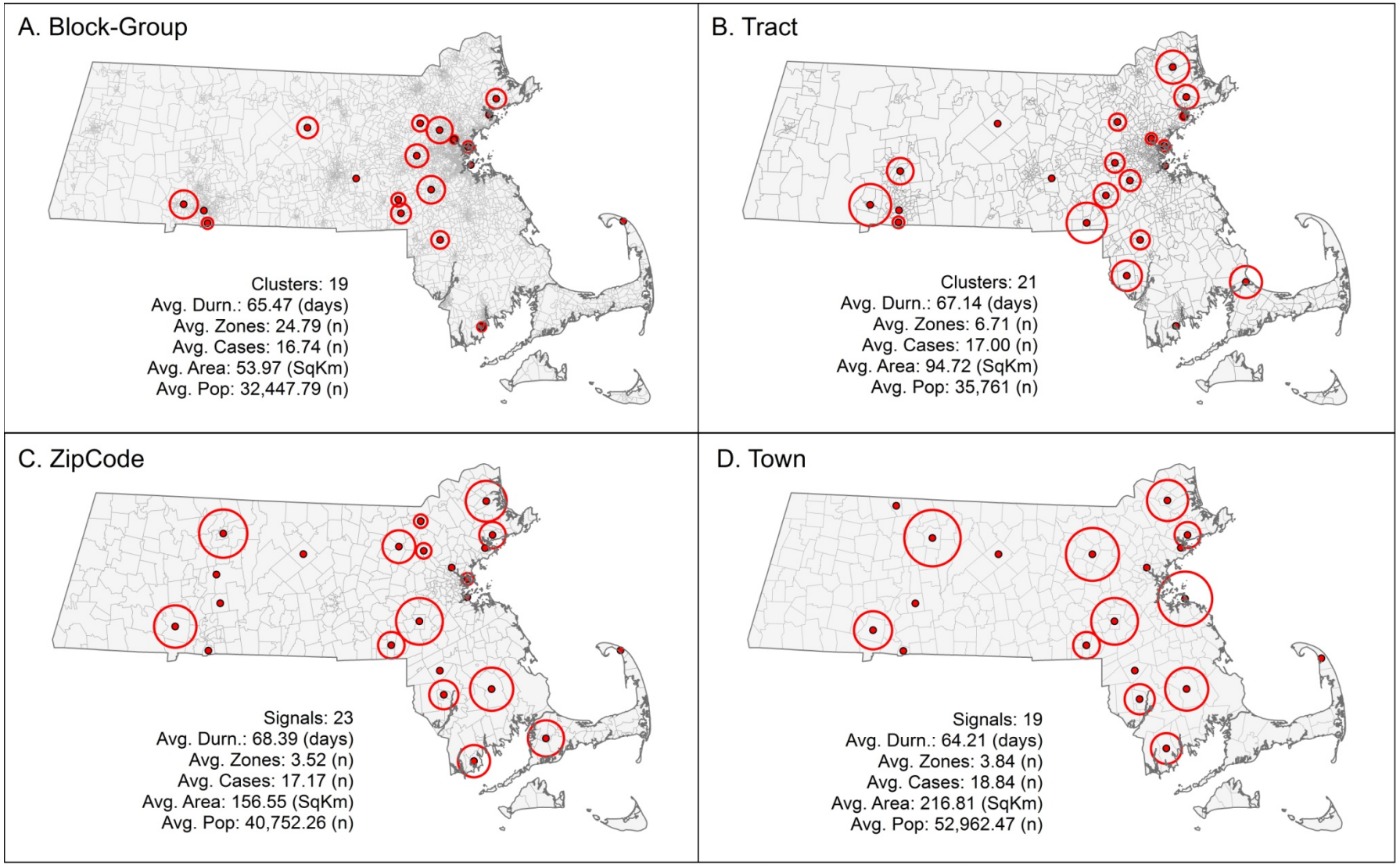
- Iterations by BGP, TRC, ZIP, TWN
- Individual maximum settings derived from known clusters
  - Eliminates oversized clusters
- Time aggregation by pertussis Infectious period (21 days)
  - Improves performance
- Software output contains indicators
  - Component zones, population, total events, start/end dates.

SaTScan cluster detection maximum limits

Parameter	Value
Max Cluster Length (days)	96
Max Cluster Radius-Census block-Group (km)	7
Max Cluster Radius-Census Tract (km)	11
Max Cluster Radius-Zip-Code (km)	12
Max Cluster Radius-City/Town (km)	14

SaTScan software cluster detection maximum limits derived from known clusters

Significant signals (p<.05) by zone type



Significant signal average indicators by zone type (p<=.05)

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

Matching Clusters & Signals

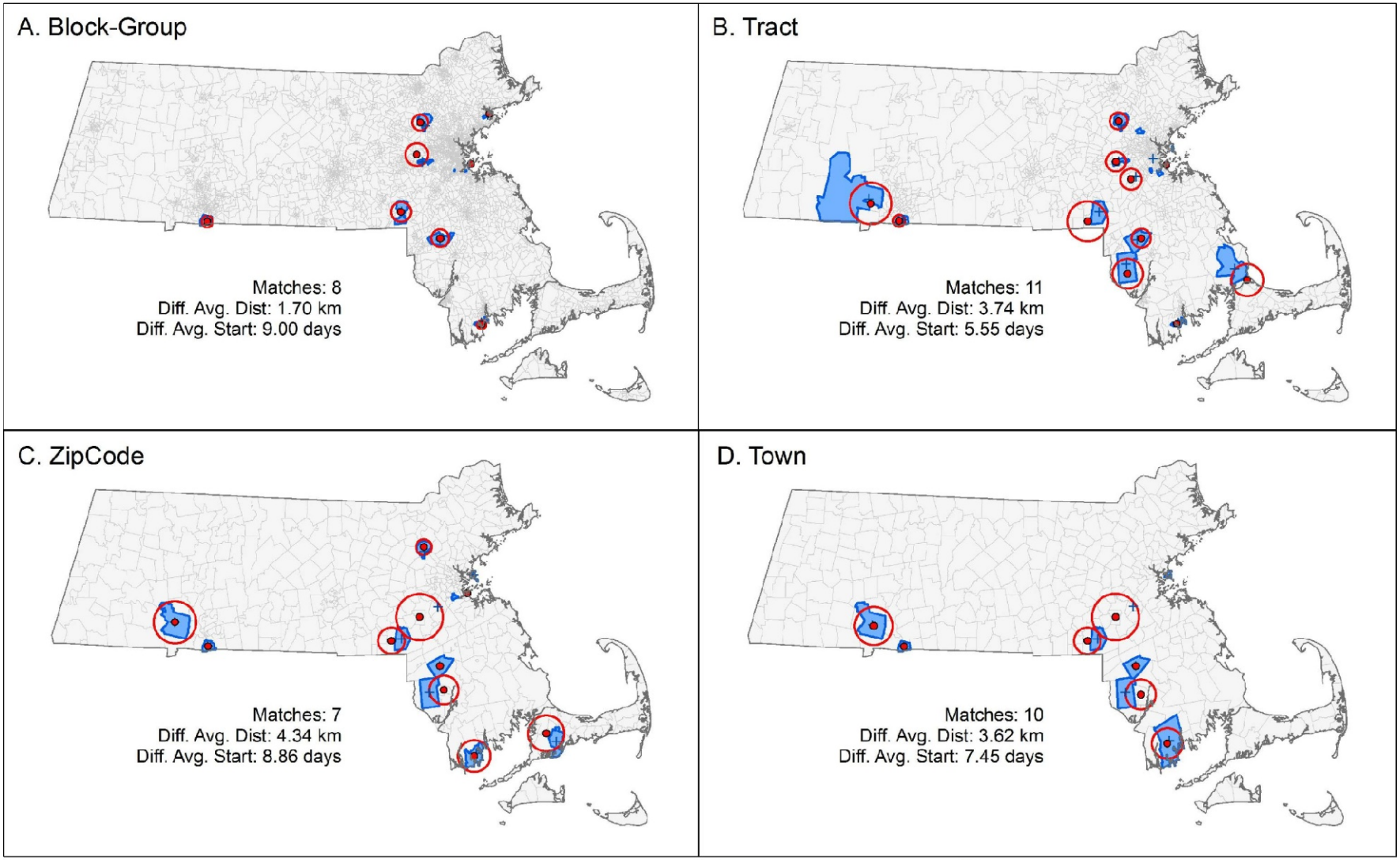
- Space-time merge using temporal window and distance between centers.

$$\exists match_i \left[\begin{array}{c} ([ClusStrt_c, ClusEnd_c] \cap [SignStrt_s, SignEnd_s]) \neq \emptyset \\ and \\ ([ClusLocn_c] \cap [SignLocn_s]) \neq \emptyset \end{array} \right]$$

  - 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



Performance Evaluation

- Sensitivity (ability to accurately detect clusters)
  - *Percent clusters detected (signals$_n$ / clusters$_n$)* <%>
- Specificity (temporal):
  - Δ in start dates *abs([ClusStrt$_i$ - SignStrt$_i$])* <days>
  - Δ in duration *abs([SignDurn$_i$ - ClusDurn$_i$])* <days>
- Specificity (spatial):
  - Δ between centers: $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ <km>
  - Δ in area: *abs([ClusArea$_i$ - SignArea$_i$])* <km^2>
  - Δ in component zones: *abs([ClusZones$_i$ - SignZones$_i$])* <n>
  - Δ in case counts: *abs([ClusCases$_i$ - SignCases$_i$])* <n>
  - Δ in population-at-risk: *abs([ClusPop$_i$ - SignPop$_i$])* <n>

Matched signals and clusters average indicator differences by zone type

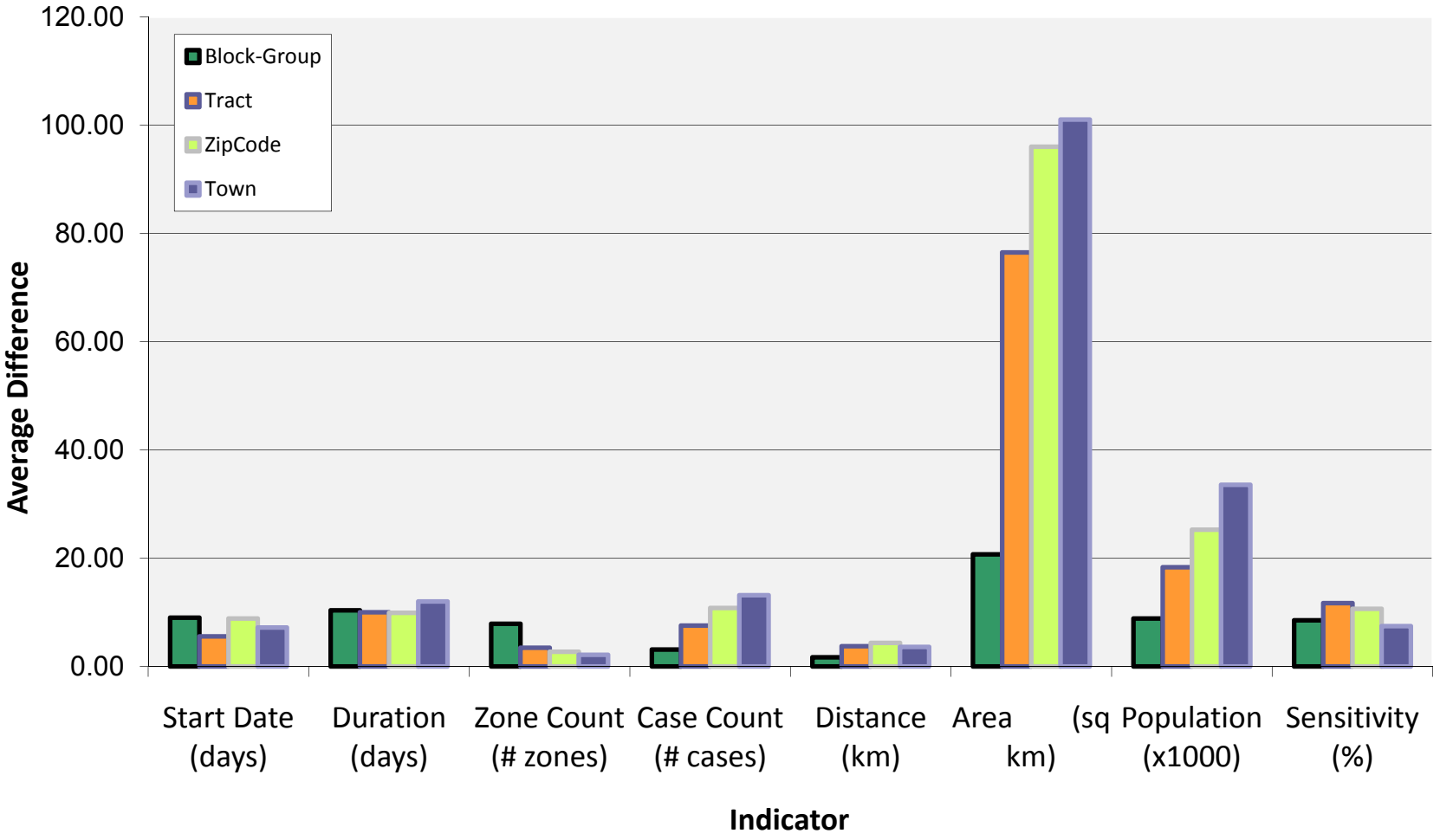
Zone Type	Total Matches	Sensitivity (% Match)	Average Start Date Difference (days)	Average Duration Difference (days)	Average Distance Difference (Km)	Average Area Difference (Sq Km)	Average Zone Count Difference	Average Case Count Difference	Average Population Difference (x1000)
Block-Group	8	**8.51**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

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