Space-Time Monitoring of Count Data for Public Health Surveillance

Arda Vanli[1], Nour Alawad[1], Rupert Giroux[2]

[1] *Department of Industrial and Manufacturing Engineering,*
Florida A&M University – Florida State University College of Engineering

[2] *Florida Department of Transportation,*
State Safety Office

ENBIS 2021 Online Spring Meeting: Data Science in Process Industries
17 – 18 May 2021

FAMU-FSU
COLLEGE OF ENGINEERING

Outline

- Introduction: Geographical anomaly detection and public health surveillance
- Proposed Methodology: Space-time CUSUM for trends in Poisson counts
- Simulation results
- Case study results

Geographical anomaly detection

- Fast statistical anomaly detection on streaming large scale data
- Anomaly: any pattern that is different from behavior that was "expected" or "normal" based on past information
  - Anomalies are time segments (for Temporal), regions (for Geographic) or vertices/edges (for Network) that have larger incidence rates than would be observed under normal conditions
- Hypothesis testing
  - H_0: Data comes from a reference pdf that is homogeneous through space
  - H_1: Inhomogeneities (clusters) exist in the data pdf
- Continuous monitoring of a system
  - Sequential measurements from system, test hypothesis, detect any anomaly as quickly as possible
  - False alarm prob, misdetection prob, reference pdf, divergence from normalcy

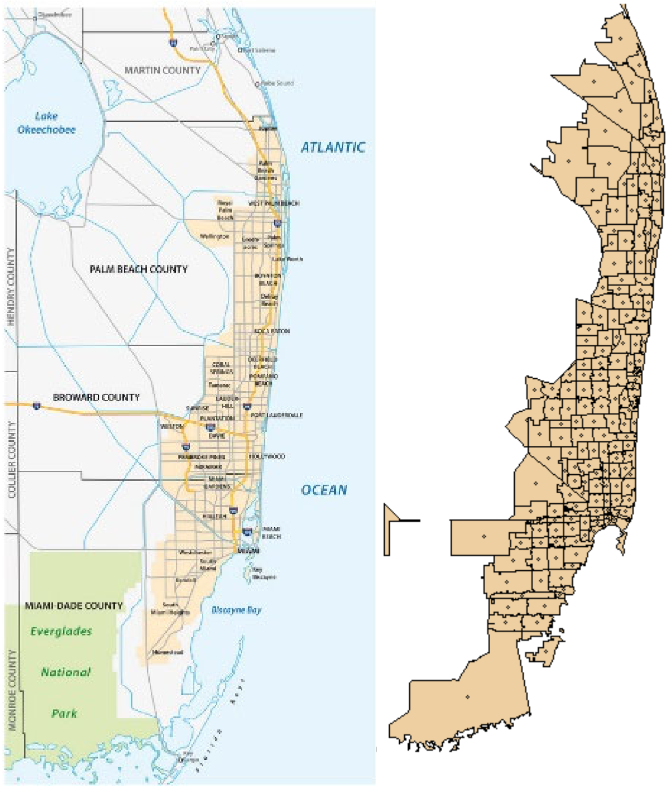
Public health surveillance

- Detect geographical clusters, or regions of anomalous activity: identify the nearby areal units with incident rates higher than expected (baseline) values
  - Baseline disease rates are estimated from historical data
  - Spatio-temporal surveillance: Sequentially take measurements of disease incidences from the map of areal units; test the hypothesis of no spatial clusters
  - Is pattern observed in today's map different from a pattern that was "expected"? Is the observed deviation from expected pattern result of some noteworthy event (e.g., Disease outbreak, contamination in waterways, crime counts in neighborhood, traffic crashes on roadways)?
- Timely and accurate detection of emerging geographical disease clusters is critical to devise effective epidemic containment and mitigation policies

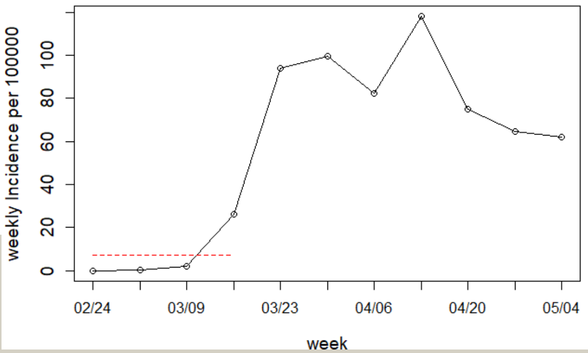
Public health surveillance: Covid-19 Outbreaks in Florida

- Weekly COVID-19 case counts observed from Feb to May 2020 in Miami, FL
  - Zip Codes are areal units
  - Expected value is found by assuming all counts are spatially and temporally homogeneous, from the first 4 weeks of data
- Identify emerging geographical clusters with higher than "expected" incidence rates
  - Determine *when* and *where* anomalously high disease rates are beginning to occur

Miami, FL, zip codes



Covid outbreaks in Miami, FL, 2020



Proposed method: Space-time CUSUM

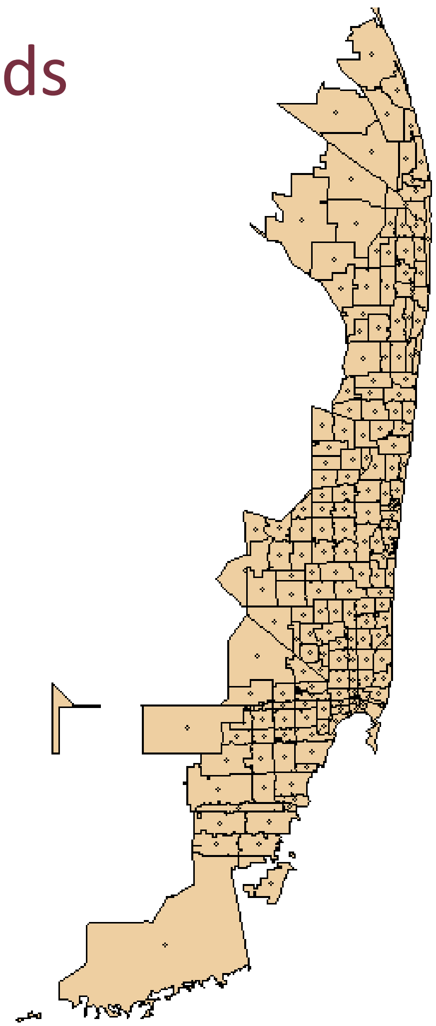
- Space-time CUSUM to detect trend-type shifts in regional Poisson count data*
  - Enumerate a set of overlapping cylinders with circular bases over a geographic area (with varying centers and radii), and a sliding interval of time (varying heights)
    - Cylinder Z with circular base centered at c, radius r and height that correspond to the time period $[\tau, \mathrm{k}]$ between outbreak onset time τ and current time k
  - Calculate the local CUSUMs for all possible cylinders, find the maximum of all local CUSUMs→ most unlikely cylinder under the null hypothesis that the rate of incidents is homogeneous over the entire space
  - Estimate of the onset of outbreak (change point)
  - Space-time CUSUM assumes a hypothesized outbreak infection rate (Sonesson, 2007). Space-time Scan uses a generalized likelihood ratio test (Kuldorff, 2001)

* Vanli, Alawad, (2020), Space-Time Surveillance of Count Data Subject to Linear Trends, *Quality and Reliability Engineering International*

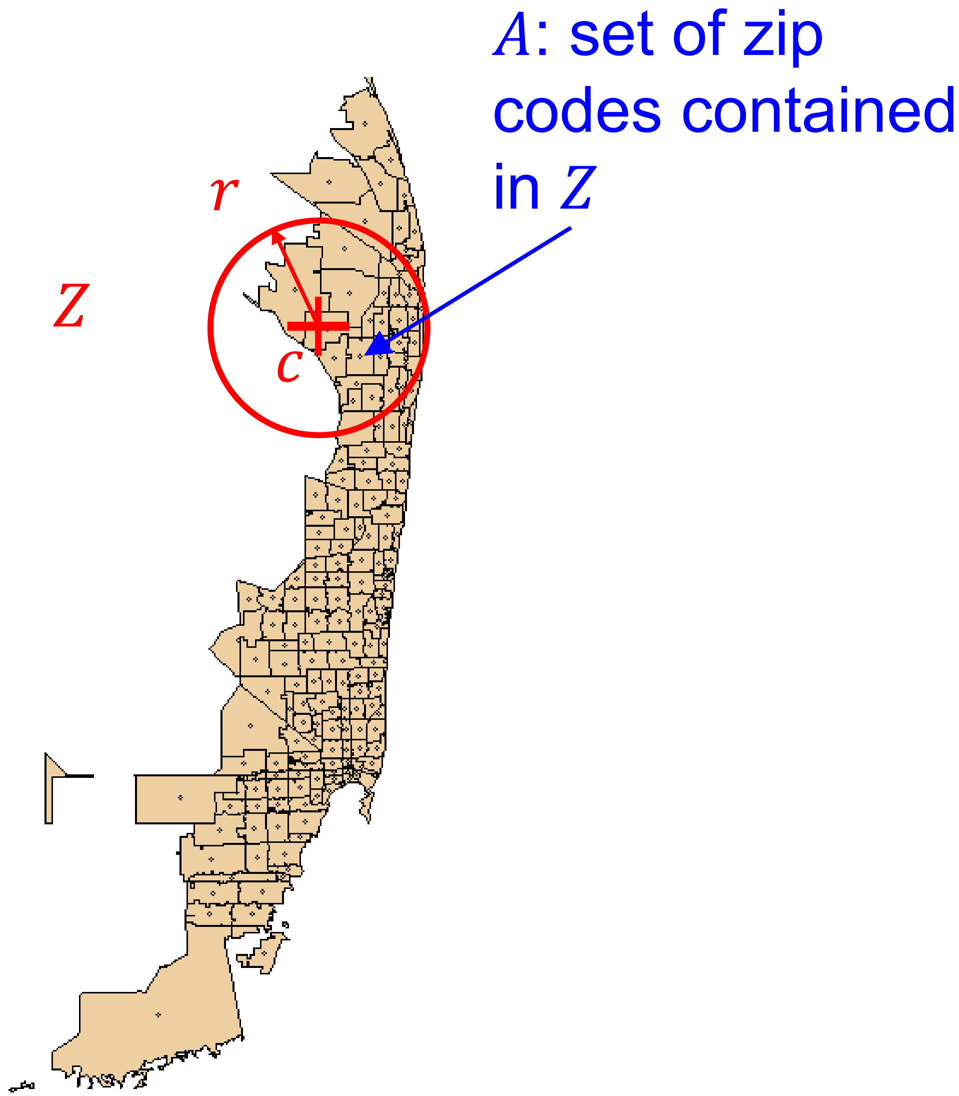
Study area: Miami, FL metro area

Areal units: Zip codes

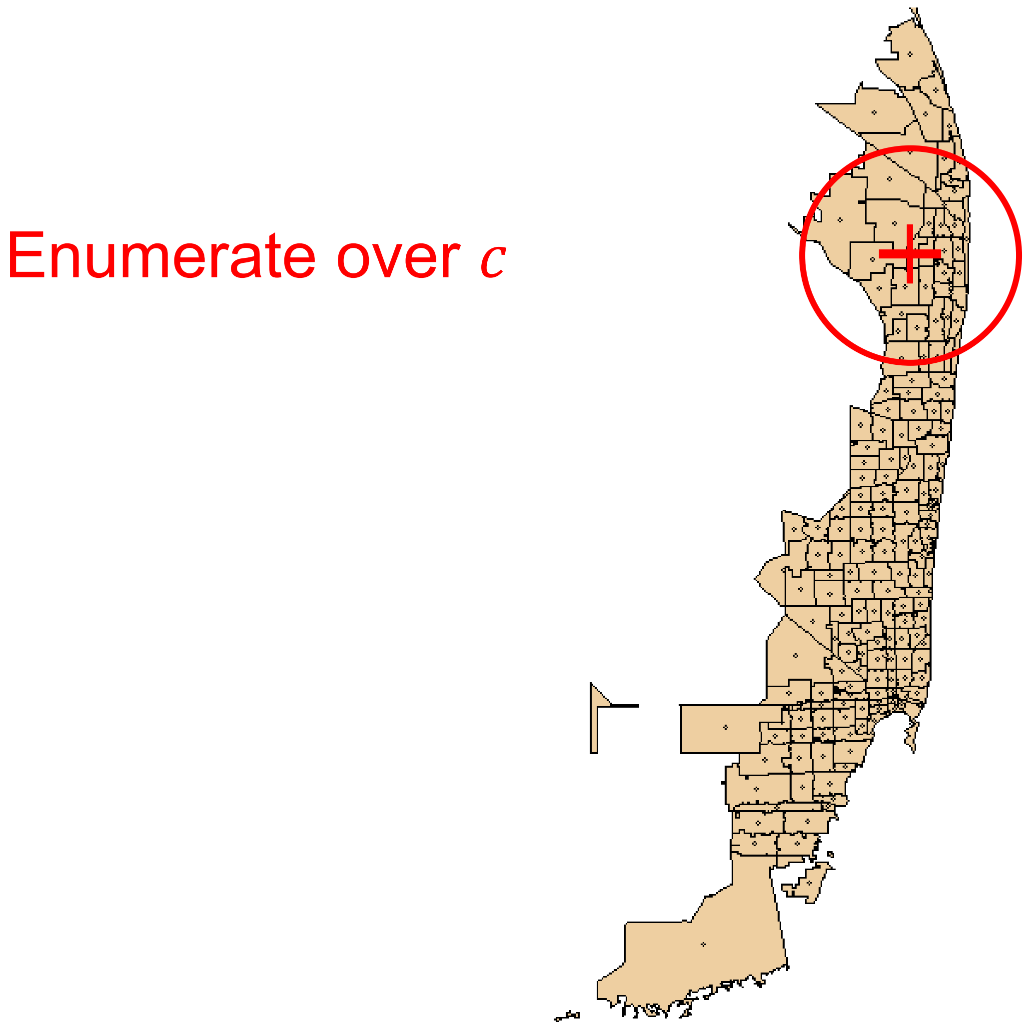
Search grid: Zip code centroids



Circular spatial boundary to aggregate counts over space

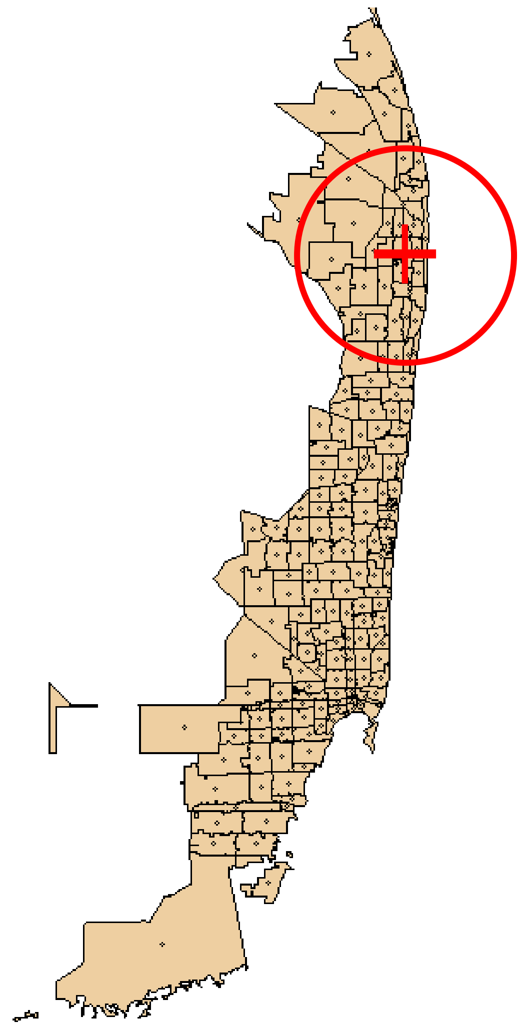


Circular spatial boundary to aggregate counts over space



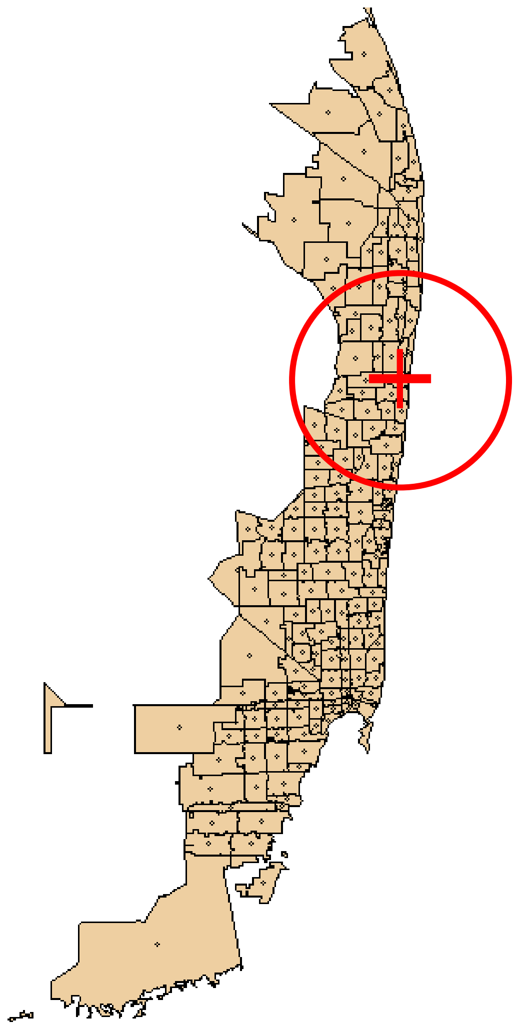
Circular spatial boundary to aggregate counts over space

Enumerate over c



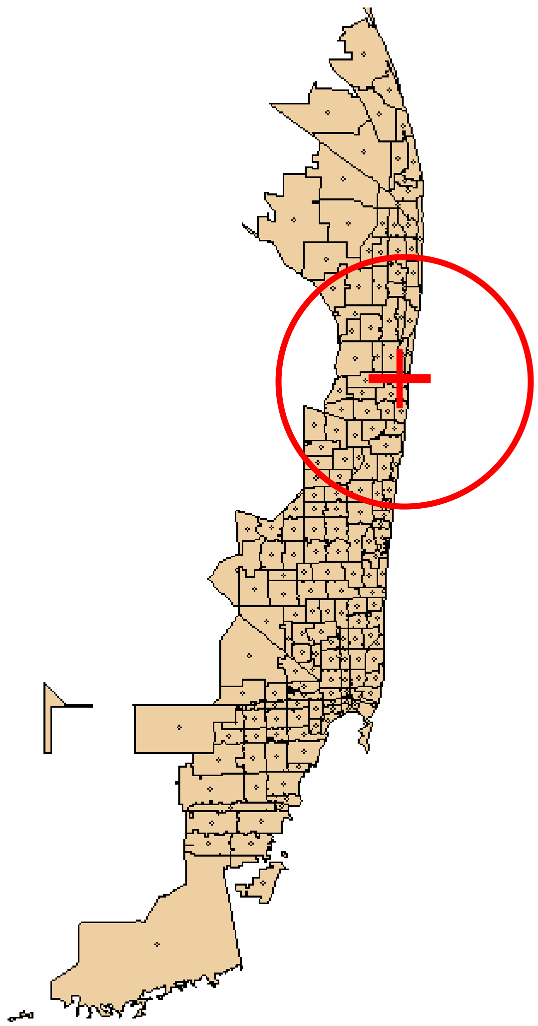
Circular spatial boundary to aggregate counts over space

Enumerate over c



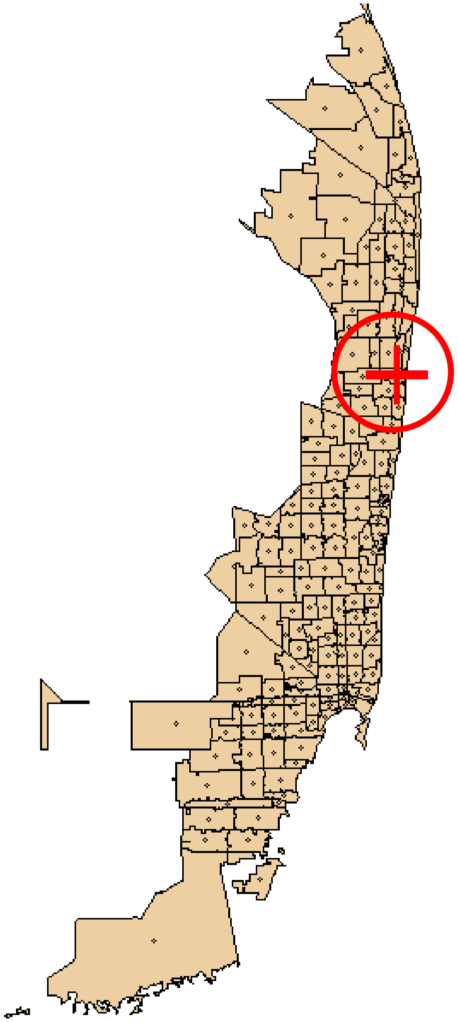
Circular spatial boundary to aggregate counts over space

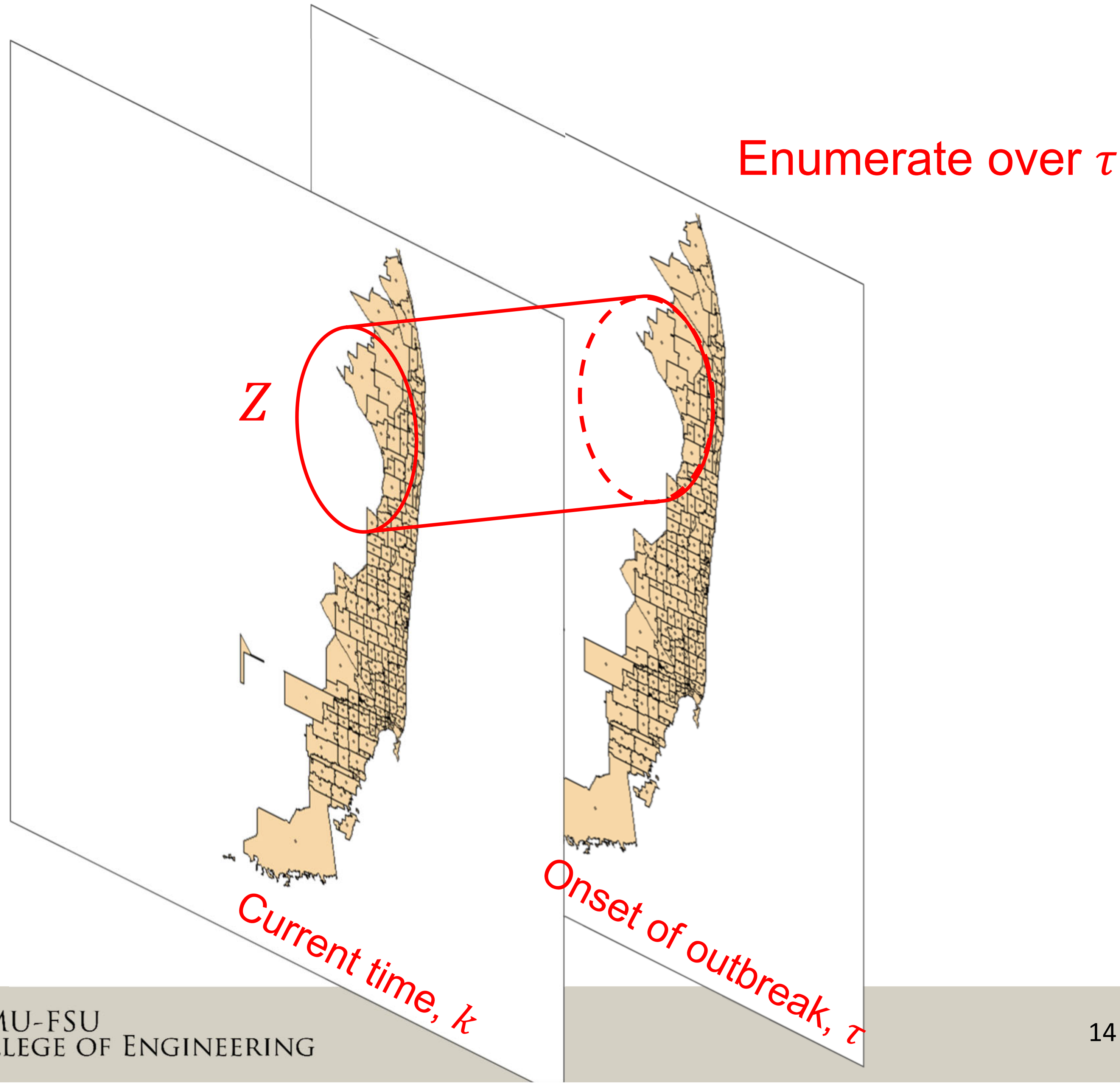
Enumerate over r

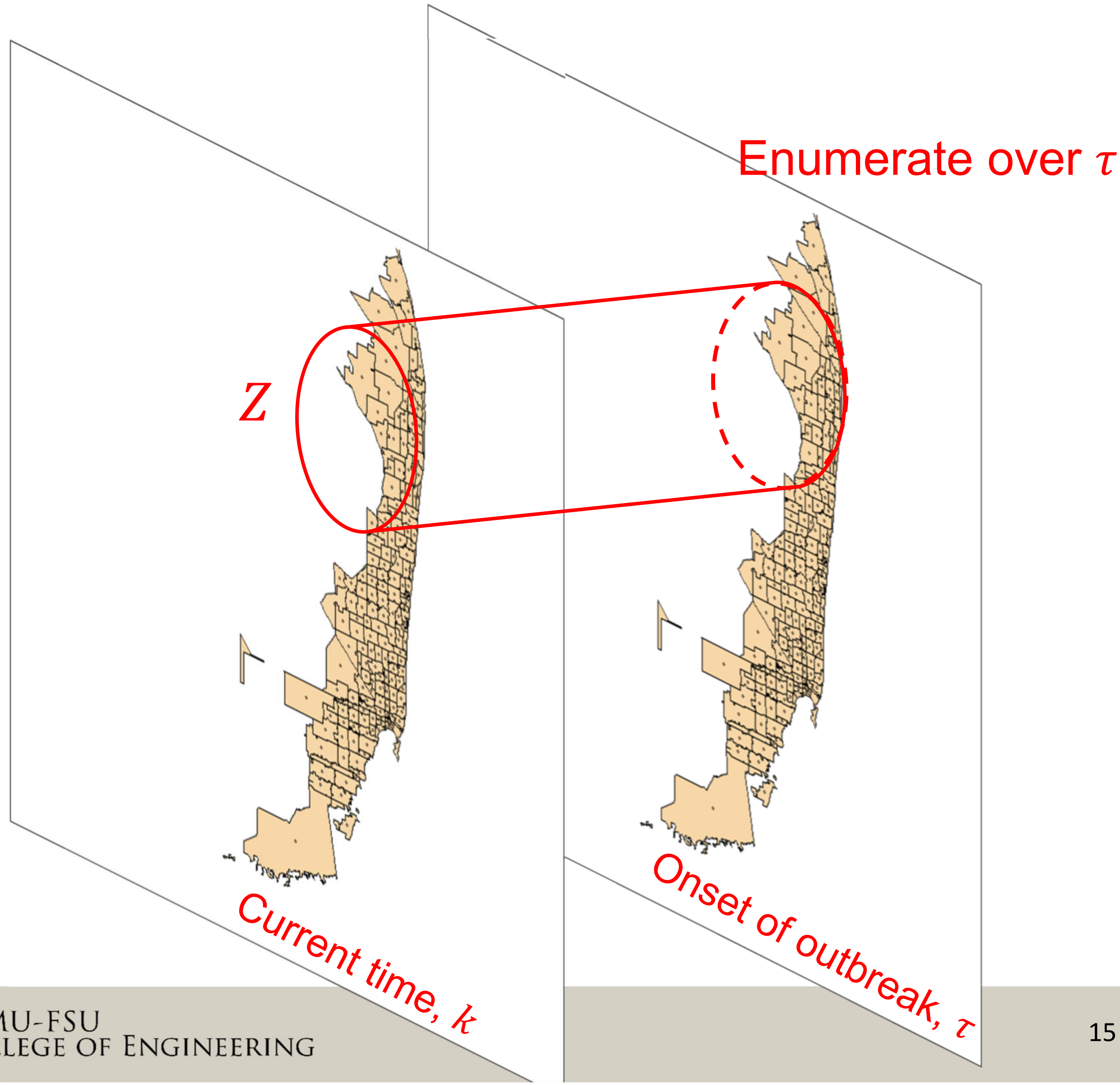


Circular spatial boundary to aggregate counts over space

Enumerate over r




Enumerate over τ
Z
Current time, k
Onset of outbreak, τ


Enumerate over τ
Z
Current time, k
Onset of outbreak, τ

Proposed method

- Count data y_{ik} at subregion i and time t follows Poisson with incidence rate μ_0. Test the spatial hypotheses:
  - H_0: infection rate is μ_0 for all sub-regions
  - H_1: infection rate of some contiguous sub-regions has shifted according to $\mu_1^*(t) = \mu_0 + \theta^* \sqrt{\mu_0}(t - \tau + 1)$, with drift rate θ^* at the change-point τ
- The log likelihood ratio (LLR) of data observed up to current time k, within the set A of subregions (cylinder Z) and time interval $[\tau, k]$

$$L(c, r, \tau, k) = \sum_{t=\tau}^{k} \log \prod_{i \in A} \frac{f(y_{it}|\mu_1^*)}{f(y_{it}|\mu_0)}$$

- Maximum of LLR over all τ to determine change-point is a local cumulative sum (CUSUM)

$$\max_{1 \le \tau \le k} L(c, r, \tau, k) \equiv T_{rc}(k) = \max \left\{ 0, T_{rc}(k-1) + \log \prod_{i \in A} \frac{f(y_{it}|\mu_1^*)}{f(y_{it}|\mu_0)} \right\}$$

Proposed method

- Maximum of the LLR over all τ, c and r to determine change-point, location and geographical size

$$\max_{r \in R} \max_{c \in C} \max_{1 \leq \tau \leq k} L(c, r, \tau, k) = \max_{r \in R} \max_{c} T_{rc}(k)$$

  – maximum of local CUSUMs, T_{rc}, defined for all r and c

- Center and size of the most likely cluster emerging at t

$$\{r^*, c^*\} = \underset{r \in R, c \in C}{\operatorname{argmax}} T_{rc}(t)$$

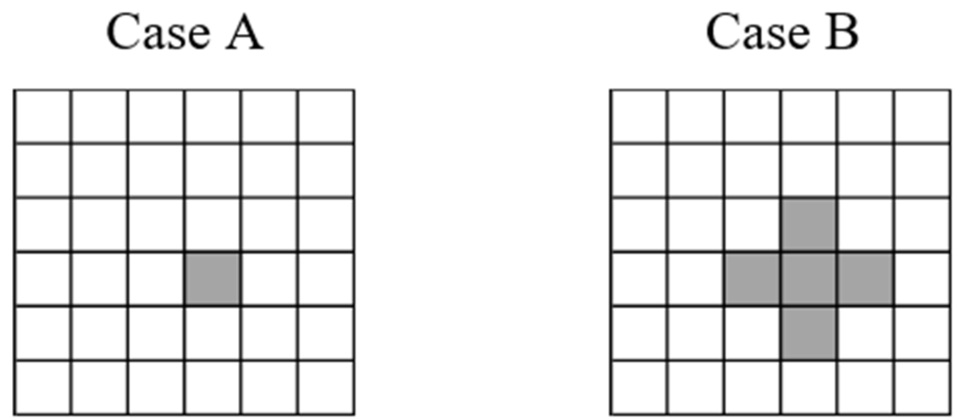
- Change-point estimate

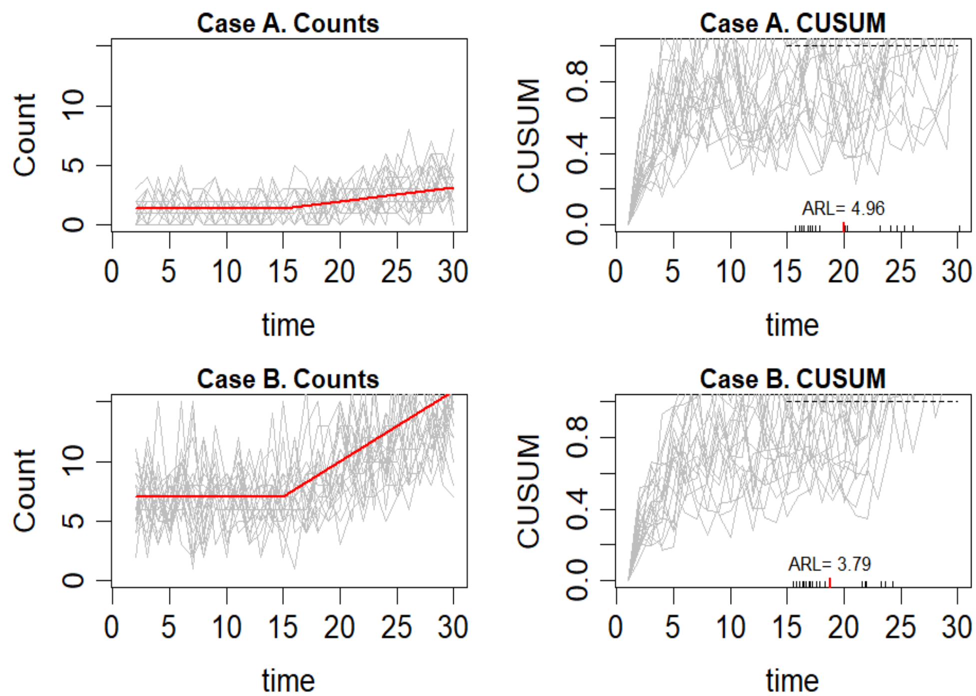
$$\hat{\tau}_{r^*, c^*} = \max_{1 \leq t \leq k} \{t | T_{r^* c^*}(t) = 0\} \text{ where}$$

- Compare to Sonesson (2007) which considered detecting step-type sustained shifts in Poisson counts

Simulation study

- Infection rate μ_0 is homogeneous spatially and it starts shifting at time t_0 with slope θ according to $\mu_1 = \mu_0 + \theta\sigma_0(t - t_0 + 1)$
  - Baseline rate $\mu_0 = 1.4$
  - slope $\theta = 0.1$
- Outbreak scales: Case A (localized) and Case B (regional)
- The set of possible radius and center values used in monitoring: $r \in \{0,1,2\}$ for $c \in \{1,2,\ldots,36\}$
- Results from 20 sample simulations where trend shift is introduced at $t_0 = 15$.

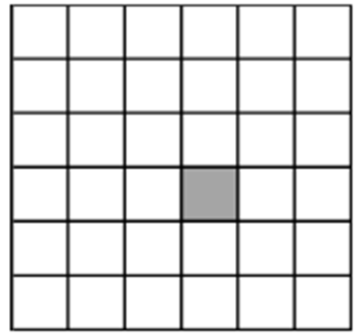




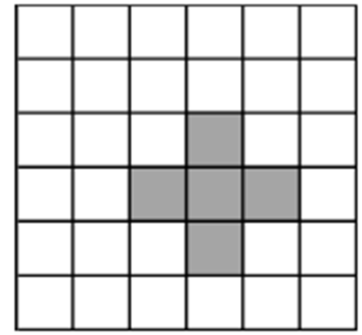
Simulation Study

- Proposed trend-type CUSUM (T-SCUSUM) was tuned to detect slopes $\theta^* = 0.10, 0.50$ and 1.00
- Sonesson (2007) step shift-type CUSUM (S-SCUSUM) was tuned to detect shifts of sizes $\delta^* = 0.81, 2.14$ and 4.00 (standard deviations)
- Methods are used to monitor outbreaks that started at time $t = 30$ with drifts $\theta = 0.10, 0.50$ and 1.00
- Simulations repeated 10,000 times

Case A



Case B



Trend and step shift type CUSUMs are designed to have approximately same ARL

	Slope	T-SCUSUM (θ^*)			S-SCUSUM (δ^*)		
Case	θ	0.1	0.5	1	0.81	2.14	4.00
A	0.1	4.31	6.61	7.87	4.25	6.28	6.72
	0.5	2.90	3.71	4.13	2.92	3.68	3.83
	1	2.38	2.82	2.86	2.29	2.72	2.79
B	0.1	3.17	4.63	5.63	3.20	4.77	5.36
	0.5	1.98	2.28	2.59	1.94	2.36	2.65
	1	1.59	1.65	1.80	1.47	1.67	1.88

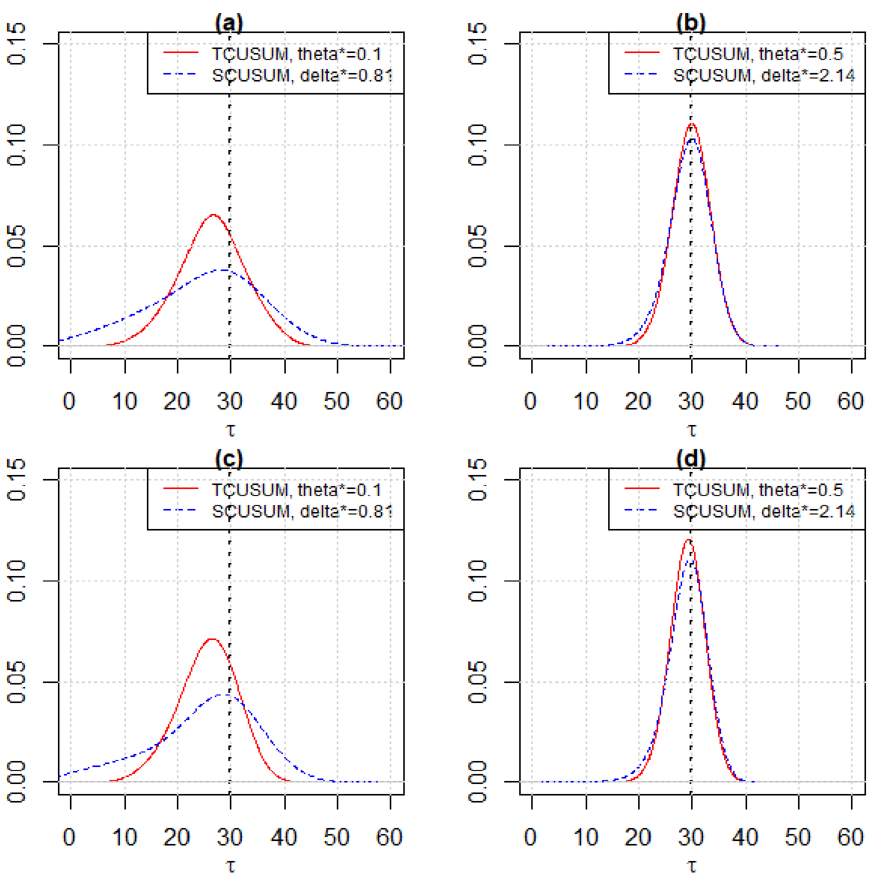
Change point estimates from T-SCUSUM and S-CUSUM

- Change point estimates $\hat{\tau}$ and the improvements obtained by the use of a trend type detector over the use of a step type detector

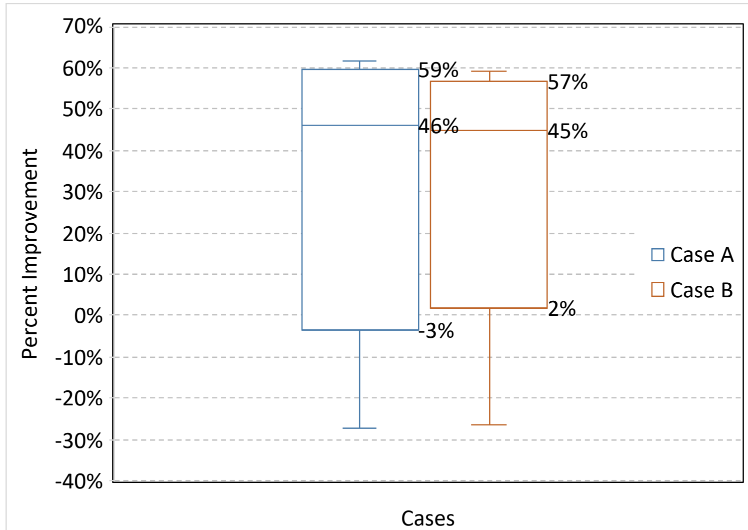
$\theta = 0.10$ $\theta = 0.50$

Case A

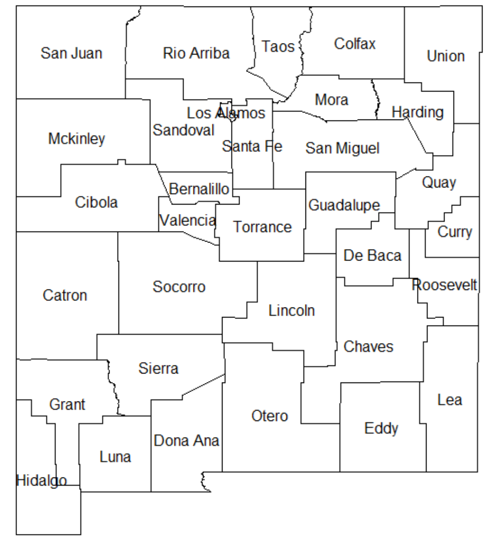
Case B



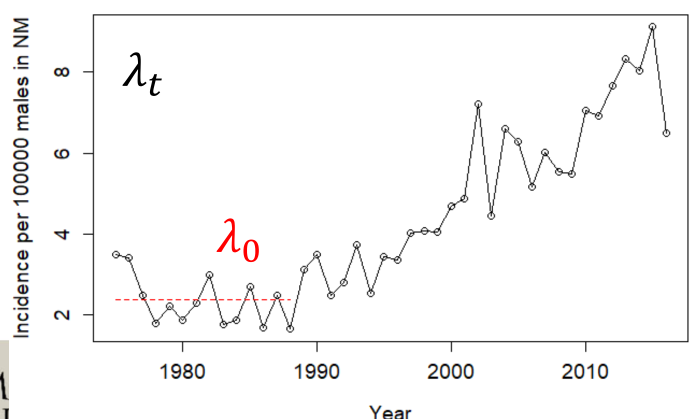
Improvement in change point $MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i - t_0)^2$ of T-SCUSUM over S-SCUSUM

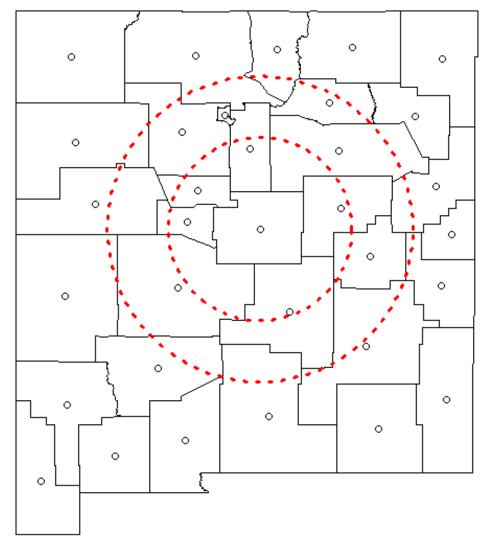


Case study 1: New Mexico thyroid cancer data



- Male thyroid cancer incidences in 32 counties of New Mexico between 1975 and 2016
- Baseline incidence rate estimated from 1975 to 1988
  - Rates of counties μ_{0it} are non-homogeneous.
  - The non-homogeneities are assumed to be due to nonhomogeneous population sizes of the subregions
$$\mu_{0it} = n_{it}\lambda_0$$
  - n_{it}: population (in 100K) in county i and year t
  - Baseline rate for entire state: λ_0(per 100,000 persons)
$$\lambda_0 = \frac{1}{14}\sum_{1975}^{1988}\lambda_t \text{ and } \lambda_t = \frac{1}{32}\sum_{i=1}^{32}\frac{y_{it}}{n_{it}}$$
- Largest scan radius includes half of the state population

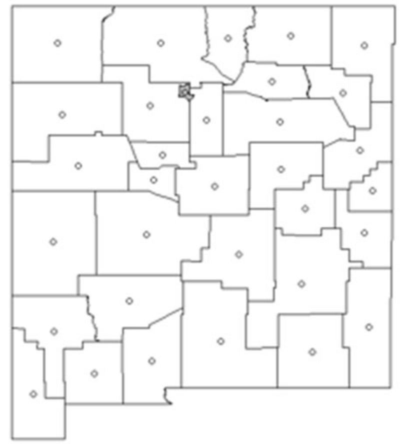




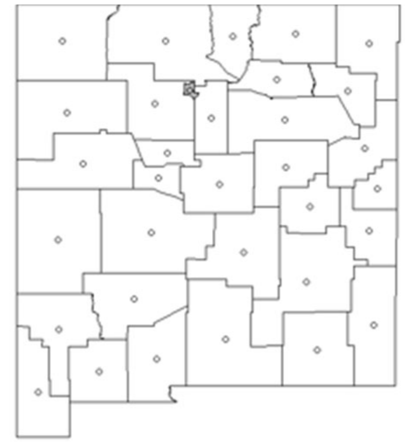
Case study 1:

- Using the T-SCUSUM (tuned for two different rates) clusters centered at Los Alamos was detected in 1994 or 1993
- Using the S-SCUSUM (tuned for two different step sizes) a cluster centered at Socorro was detected in 1995 and a cluster centered at Bernalillo was detected in 1995
- More consistent clusters are identified with trend-type detectors than step-type detectors

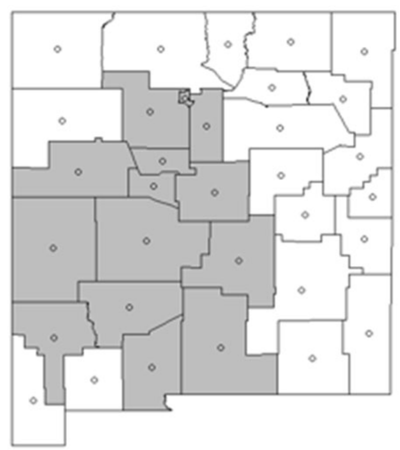
Cluster identified in 1994 with T-SCUSUM and $\theta^* = 0.25$



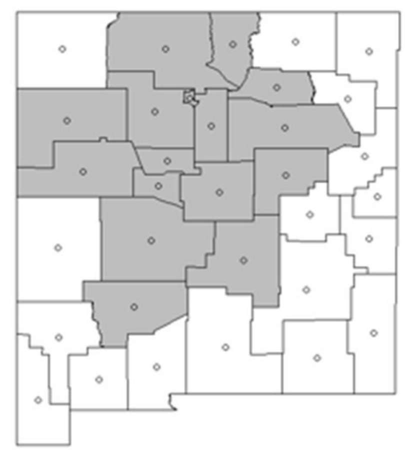
Cluster identified in 1993 with T-SCUSUM and $\theta^* = 0.5$



Cluster identified in 1994 with S-SCUSUM and $\delta^* = 0.25$



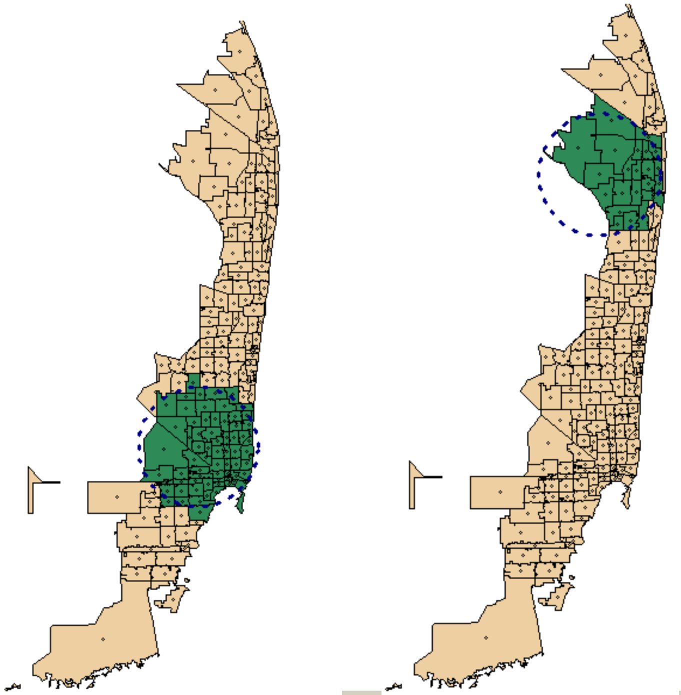
Cluster identified in 1994 with S-SCUSUM and $\delta^* = 0.5$

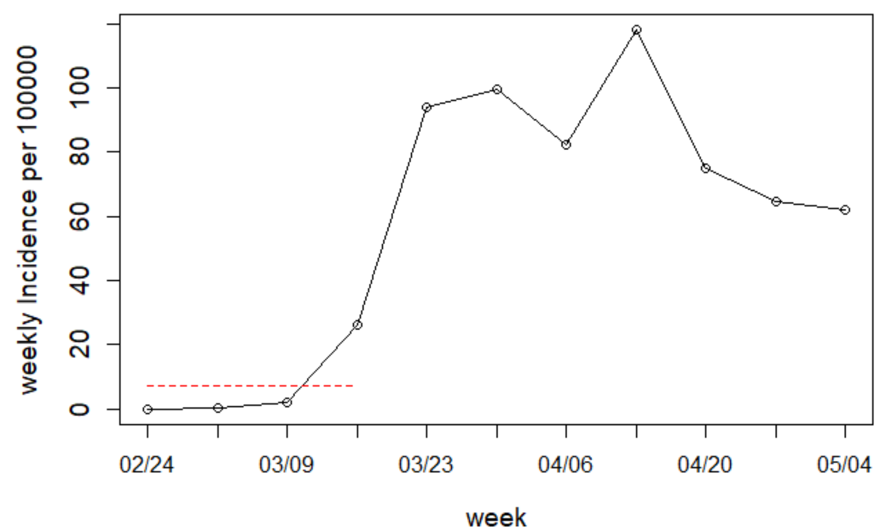


Case Study 2: Covid 19 outbreaks in Miami, FL

- Weekly COVID-19 case counts observed from Feb to May 2020 in Miami, FL
  - Space-time CUSUM with trend shift was implemented for case counts observed weekly in zip codes
  - Two outbreaks are detected at two geographically distinct clusters

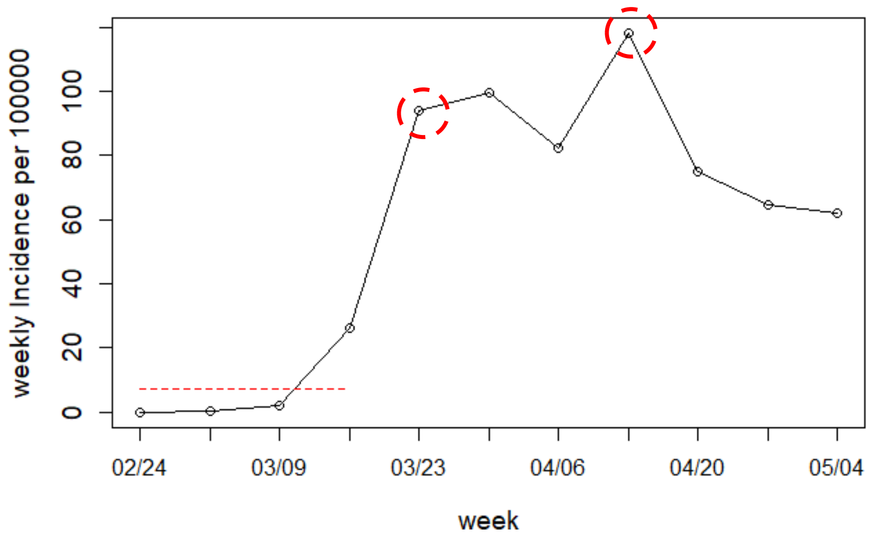
week 5 week 8

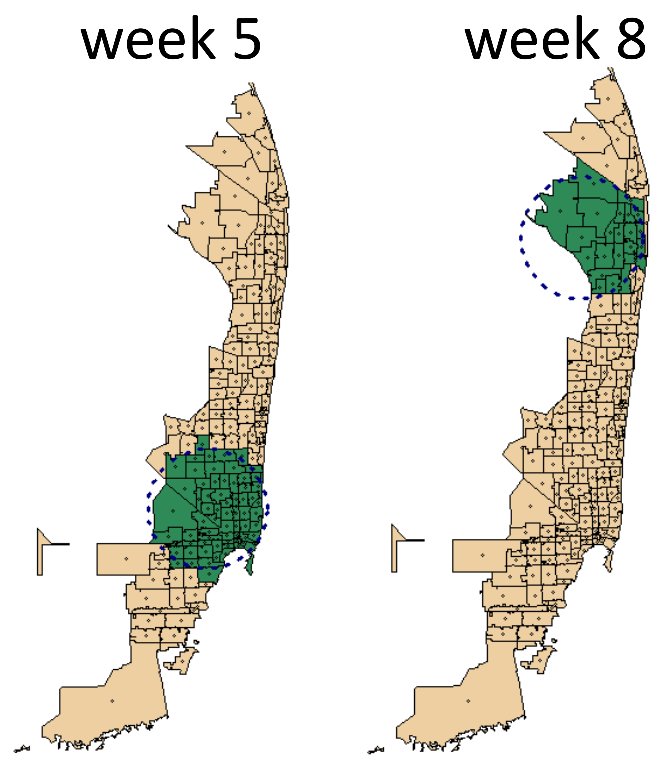


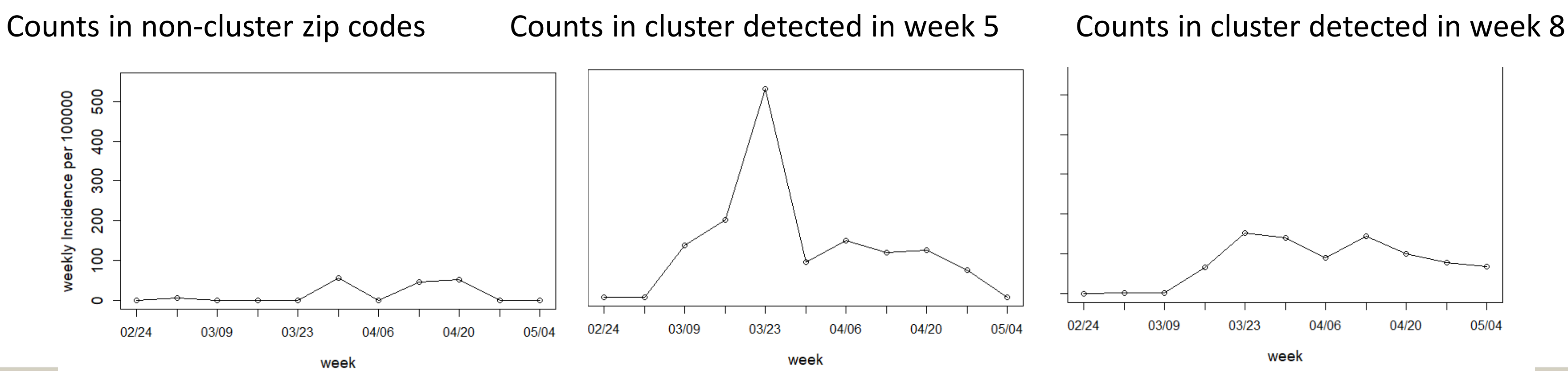


Case Study 2:

- The identified clusters have distinctly different count trajectories over time than the baseline







Conclusions

- Simulation study:
  - space-time CUSUM monitoring tuned for trend-type shifts can significantly outperform the counterparts tuned for sustained shifts in terms of the change-point estimation accuracy (MSE)
  - a practical impact: accurate change-point estimates would enable health professionals to more accurately identify and isolate the disease emergence location and time and to devise more effective epidemic containment and mitigation policies
- Case studies:
  - Thyroid cancer data: trend-type space-time CUSUM gives more consistent cluster estimates regardless of tuning
  - Miami COVID data: high resolution monitoring for zip code cases allows taking more community focused containment measures
  - Identified clusters can be useful to identify gatherings with inadequate social distancing and implement targeted community testing. E.g., New York City's Health Department used surveillance to detect COVID-19 percent test positivity clusters and formulate containment solutions (Greene et al., 2020)

Questions?

Nour Alawad

Email: na18x@my.fsu.edu

Arda Vanli

Email: oavanli@eng.famu.fsu.edu