ISDS Webinar:
Spatial Scanning Tips and Tricks for Practical Outbreak Detection

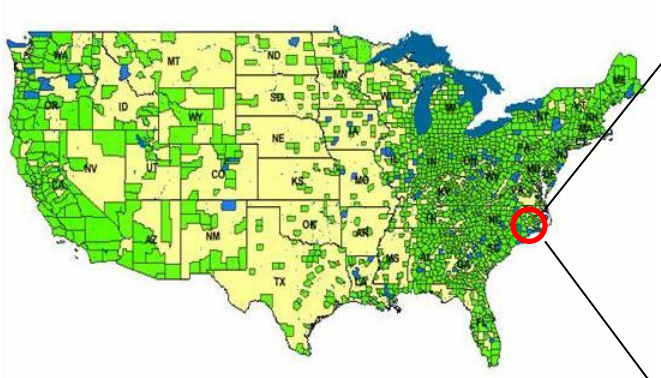
Daniel B. Neill, Ph.D.
Event and Pattern Detection Laboratory
Carnegie Mellon University
E-mail: neill@cs.cmu.edu

We gratefully acknowledge funding support from the National Science Foundation, grants IIS-0916345, IIS-0911032, and IIS-0953330.

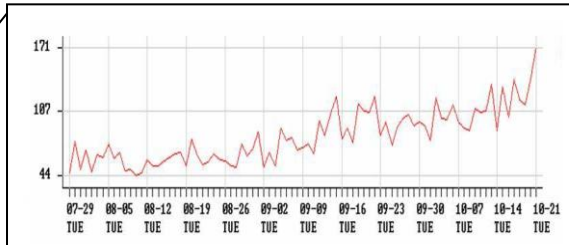
What is spatial scan?

- Spatial scan ≠ SaTScan (or, for that matter, any other single “out of the box” solution).
- Not a single method: a collection of many related methods for **spatial event detection**.
- Original spatial scan statistic by Kulldorff; many variants and extensions developed by research community over the last ~15 years.
- Different variants work better (or worse) in different circumstances → need to think carefully about which ones to use.

Spatial event detection



Spatial time series data from spatial locations s_i (e.g. zip codes)



Time series of counts $c_{i,m}{}^t$ for each location s_i for each data stream D_m.

Outbreak detection

D_1 = respiratory ED

D_2 = constitutional ED

D_3 = OTC cough/cold

D_4 = OTC anti-fever

(etc.)

Goals of detection task: **detect** any emerging disease outbreaks, **pinpoint** the affected spatial area, and **characterize** the type of event.

Informally, we want to know:

Is there anything happening?

If so, **what** and **where**?

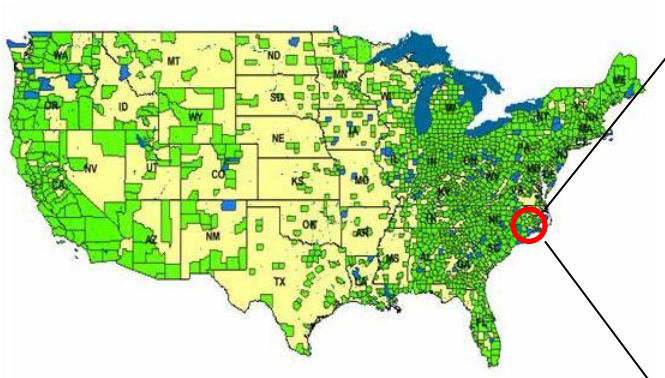
Formally, we distinguish between:

Null hypothesis $\mathbf{H_0}$ (no events)

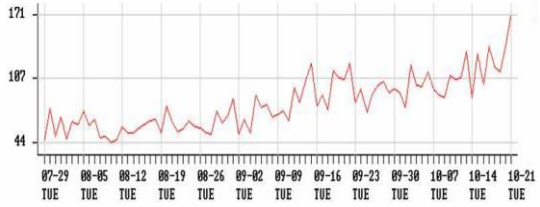
Set of alternative hypotheses $\mathbf{H_1(S, E_k)}$ = event of type E_k in spatial region S.

(Spatial region = set of "nearby" locations, often constrain shape/size)

Spatial event detection



Spatial time series data from spatial locations s_i (e.g. zip codes)



Time series of counts $c_{i,m}^t$ for each location s_i for each data stream D_m.

Outbreak detection

D_1 = respiratory ED

D_2 = constitutional ED

D_3 = OTC cough/cold

D_4 = OTC anti-fever

(etc.)

Goals of detection task: **detect** any emerging disease outbreaks, **pinpoint** the affected spatial area, and **characterize** the type of event.

Simplifying assumptions:

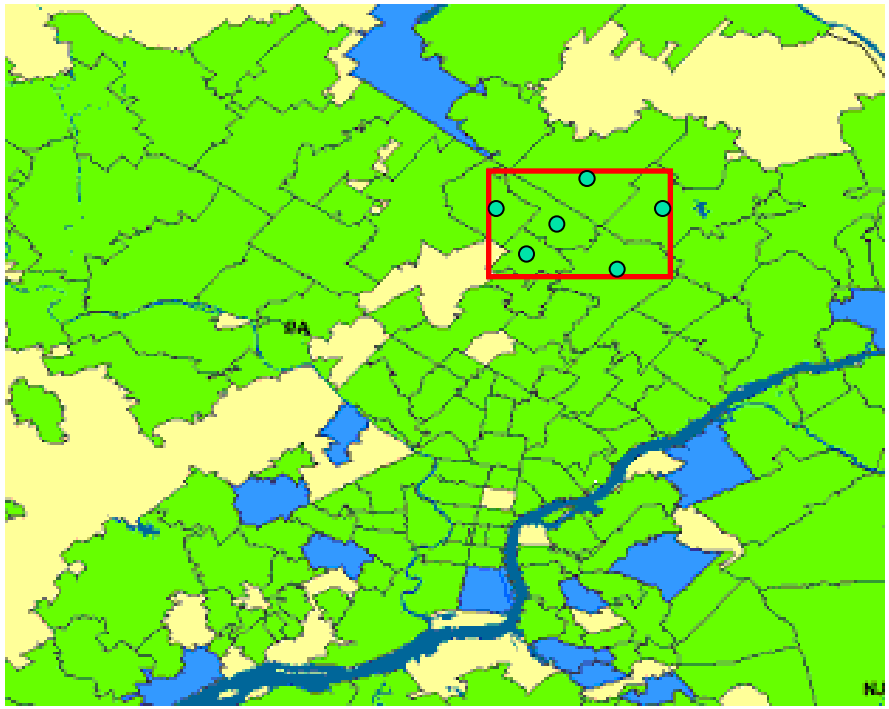
Single data stream → Consider counts c_i^t.

Single event type → Testing $H_1(S)$, "Counts in region S are significantly higher than expected."

Typically many more assumptions, e.g. counts are Poisson distributed, uniform increase in risk, …

The spatial scan statistic

(Kulldorff, 1997)

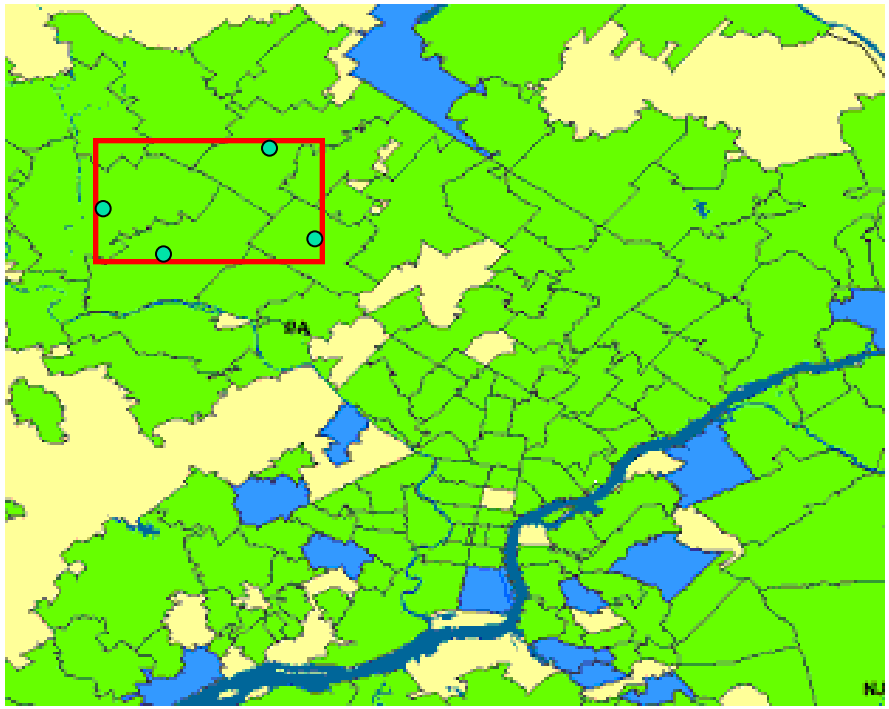


Rather than monitoring individual locations, we examine groups of locations.

Imagine moving a spatial window around the monitored area, allowing the size and shape of the window to vary.

The spatial scan statistic

(Kulldorff, 1997)

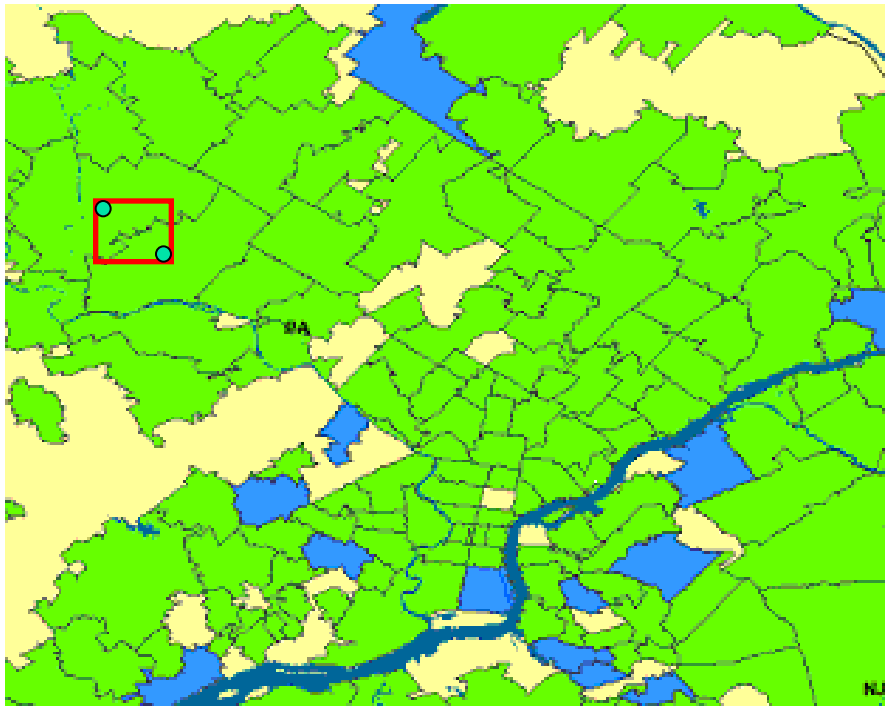


Rather than monitoring individual locations, we examine groups of locations.

Imagine moving a spatial window around the monitored area, allowing the size and shape of the window to vary.

The spatial scan statistic

(Kulldorff, 1997)

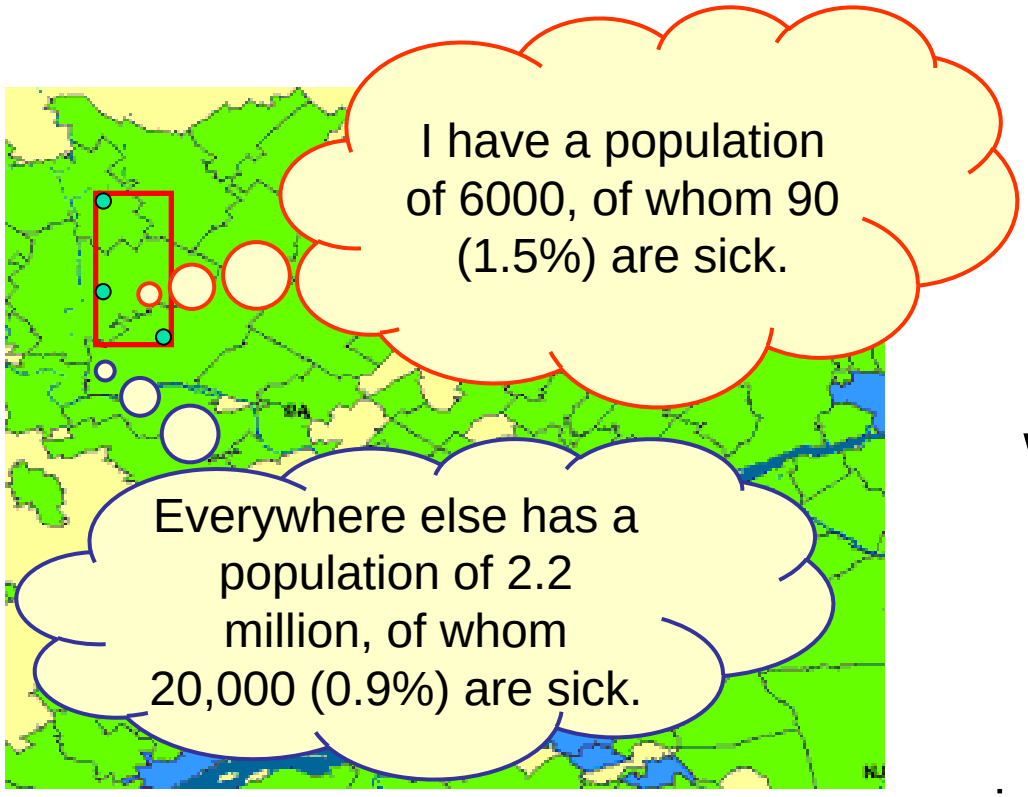


Rather than monitoring individual locations, we examine groups of locations.

Imagine moving a spatial window around the monitored area, allowing the size and shape of the window to vary.

The spatial scan statistic

(Kulldorff, 1997)



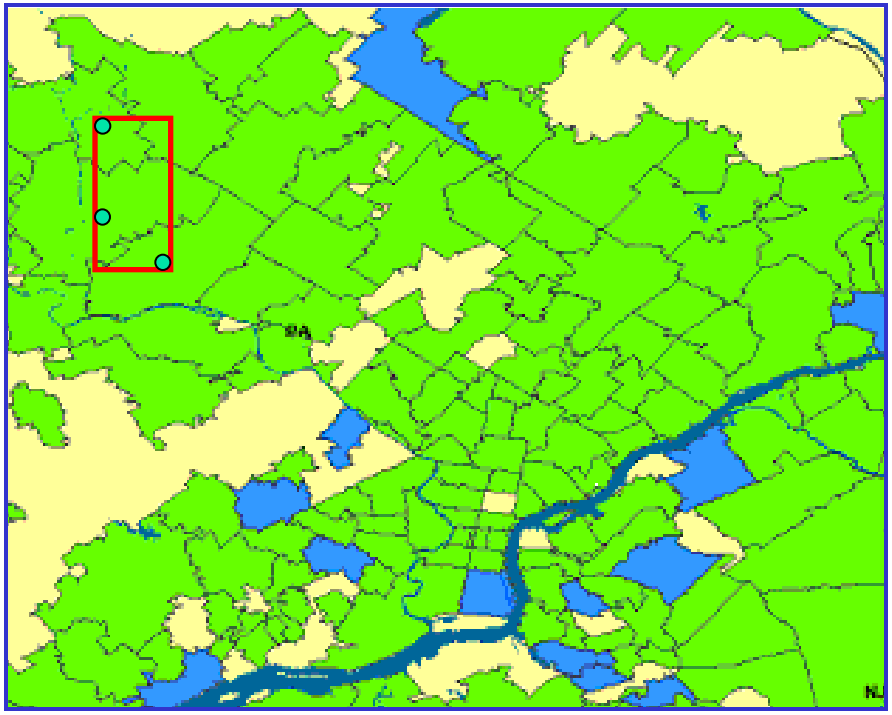
Rather than monitoring individual locations, we examine groups of locations.

Imagine moving a spatial window around the monitored area, allowing the size and shape of the window to vary.

Is there any position of the window such that the points inside form a significant cluster?

We compute a **score** for each spatial region, and then test whether the highest scoring regions are significant.

Finding the most significant regions



- Define <u>models</u>:
  - of the null hypothesis H_0: no events.
  - of the alternative hypotheses $H_1(S)$: event in region S.

c_i = **count** for location s_i (e.g. number of disease cases)

b_i = **baseline** for location s_i (e.g. population at-risk, or expected count computed from historical data)

q = **risk** (expected ratio of count to baseline)

<u>Kulldorff's model</u>

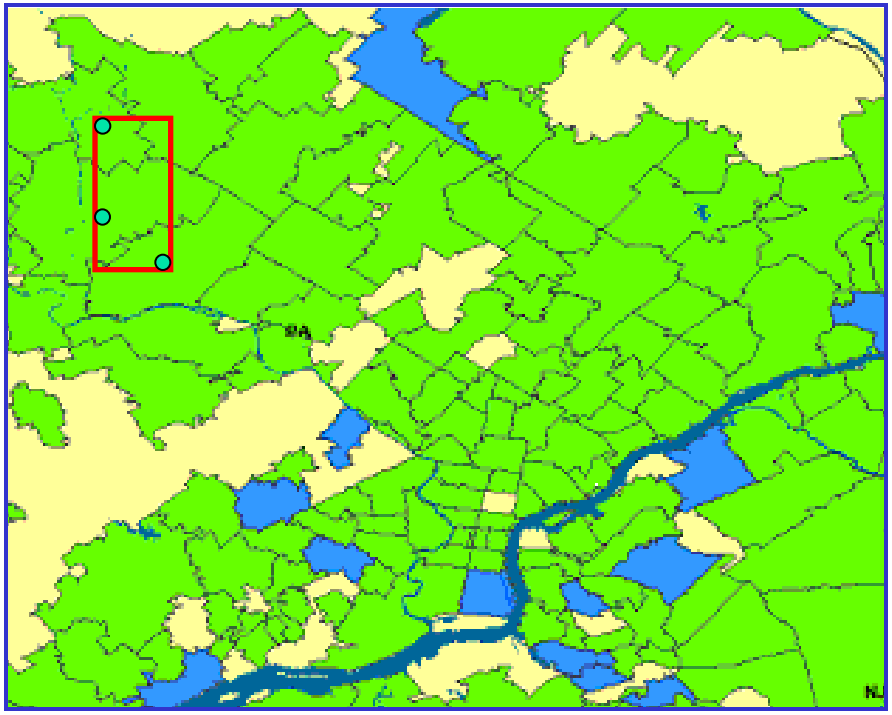
$c_i \sim \text{Poisson}(qb_i)$

H_0: $q = q_{all}$ everywhere

$H_1(S)$: $q = q_{in}$ inside S,
$q = q_{out}$ outside,
$q_{in} > q_{out}$.

Finding the most significant regions

- Define <u>models</u>:
  - of the null hypothesis H_0: no events.
  - of the alternative hypotheses $H_1(S)$: event in region S.



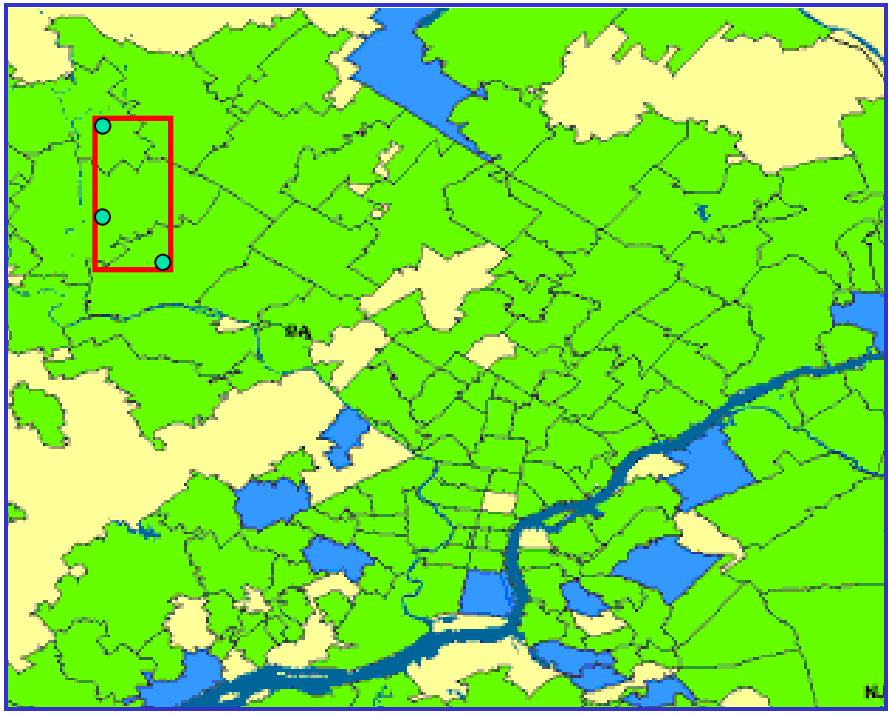
<u>Kulldorff's model</u>

$c_i \sim \text{Poisson}(qb_i)$

H_0: $q = q_{all}$ everywhere

$H_1(S)$: $q = q_{in}$ inside S,
$q = q_{out}$ outside,
$q_{in} > q_{out}$.

Finding the most significant regions



- Define <u>models</u>:
  - of the null hypothesis H_0: no events.
  - of the alternative hypotheses $H_1(S)$: event in region S.
- Derive a <u>score function</u>:
  - Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}$$

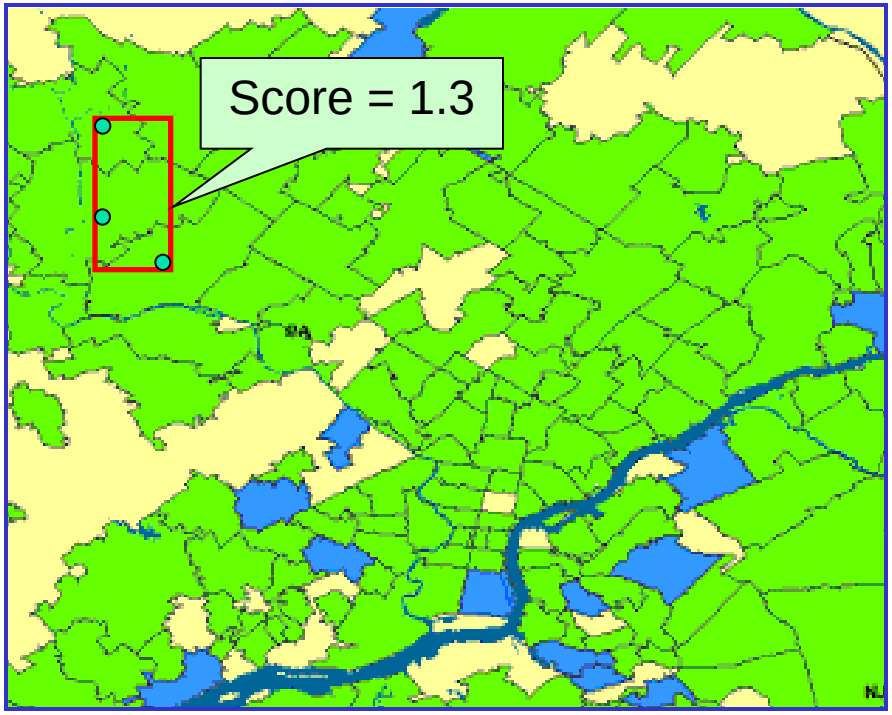
<u>Kulldorff's model</u>

$c_i \sim \text{Poisson}(qb_i)$

H_0: $q = q_{all}$ everywhere

$H_1(S)$: $q = q_{in}$ inside S,
$q = q_{out}$ outside,
$q_{in} > q_{out}$.

Finding the most significant regions



- Define <u>models</u>:
  - of the null hypothesis H_0: no events.
  - of the alternative hypotheses $H_1(S)$: event in region S.
- Derive a <u>score function</u>:
  - Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}$$

$$F(S) = \left(\frac{C}{B}\right)^{C} \left(\frac{C_{tot} - C}{B_{tot} - B}\right)^{C_{tot} - C} \left(\frac{C_{tot}}{B_{tot}}\right)^{-C_{tot}}$$

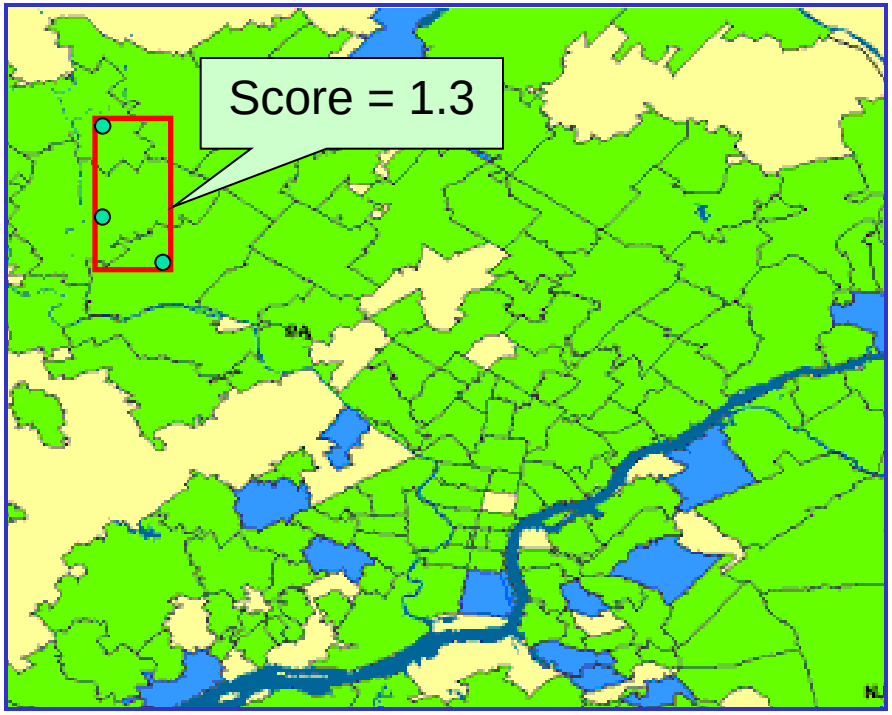
<u>Kulldorff's model</u>

$c_i \sim \text{Poisson}(qb_i)$

H_0: $q = q_{all}$ everywhere

$H_1(S)$: $q = q_{in}$ inside S,
$q = q_{out}$ outside,
$q_{in} > q_{out}$.

Finding the most significant regions



- Define <u>models</u>:
  - of the null hypothesis H_0: no events.
  - of the alternative hypotheses $H_1(S)$: event in region S.
- Derive a <u>score function</u>:
  - Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}$$

Total count and baseline of region S

Total count and baseline of search area

$$F(S) = \left(\frac{C}{B}\right)^{C} \left(\frac{C_{tot} - C}{B_{tot} - B}\right)^{C_{tot} - C} \left(\frac{C_{tot}}{B_{tot}}\right)^{-C_{tot}}$$

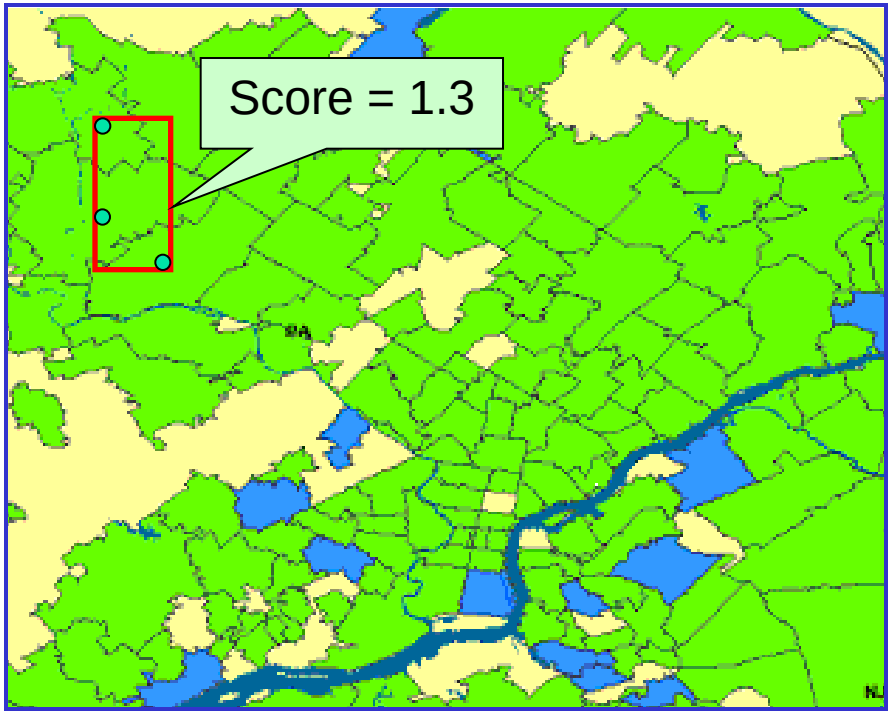
<u>Kulldorff's model</u>

$c_i \sim \text{Poisson}(qb_i)$

H_0: $q = q_{all}$ everywhere

$H_1(S)$: $q = q_{in}$ inside S,
$q = q_{out}$ outside,
$q_{in} > q_{out}$.

Finding the most significant regions



- Define <u>models</u>:
  - of the null hypothesis H_0: no events.
  - of the alternative hypotheses $H_1(S)$: event in region S.
- Derive a <u>score function</u>:
  - Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}$$

$$F(S) = \left(\frac{C}{B}\right)^{C} \left(\frac{C_{tot} - C}{B_{tot} - B}\right)^{C_{tot} - C} \left(\frac{C_{tot}}{B_{tot}}\right)^{-C_{tot}}$$

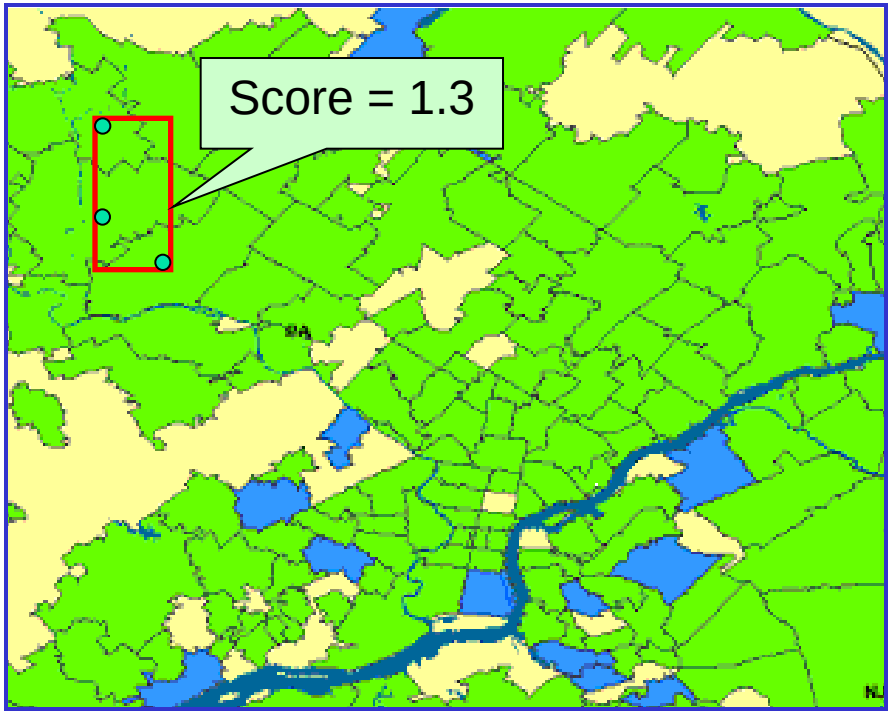
<u>Kulldorff's model</u>

$c_i \sim \text{Poisson}(qb_i)$

H_0: $q = q_{all}$ everywhere

$H_1(S)$: $q = q_{in}$ inside S,
$q = q_{out}$ outside,
$q_{in} > q_{out}$.

Finding the most significant regions



- Define <u>models</u>:
  - of the null hypothesis H_0: no events.
  - of the alternative hypotheses $H_1(S)$: event in region S.
- Derive a <u>score function</u>:
  - Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}$$

  - To find the most significant regions:

$$S^* = \arg\max_S F(S)$$

$$F(S) = \left(\frac{C}{B}\right)^{C} \left(\frac{C_{tot} - C}{B_{tot} - B}\right)^{C_{tot} - C} \left(\frac{C_{tot}}{B_{tot}}\right)^{-C_{tot}}$$

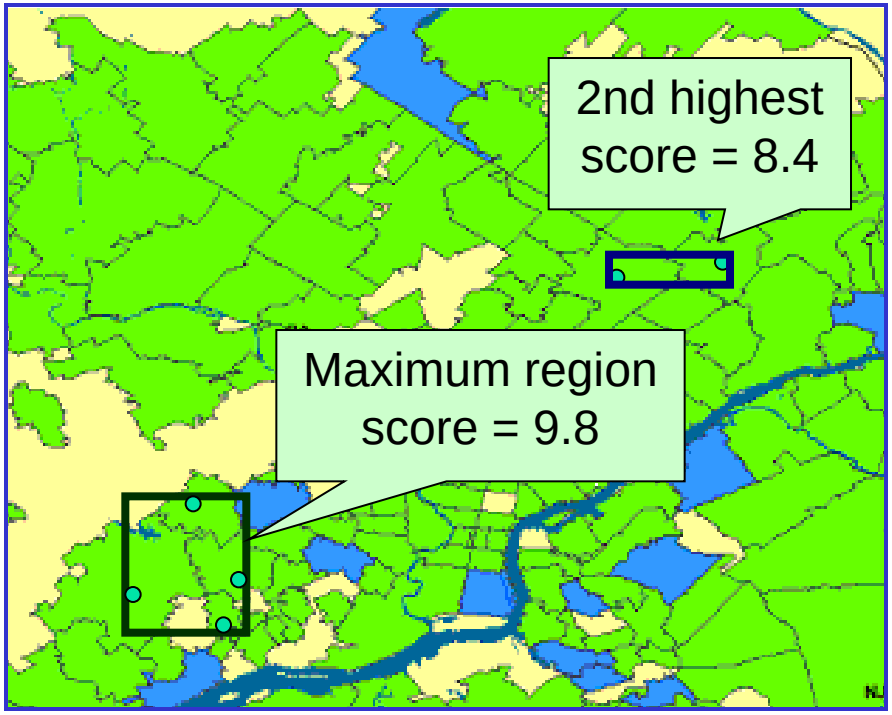
<u>Kulldorff's model</u>

$c_i \sim \text{Poisson}(qb_i)$

H_0: $q = q_{all}$ everywhere

$H_1(S)$: $q = q_{in}$ inside S,
$q = q_{out}$ outside,
$q_{in} > q_{out}$.

Finding the most significant regions



- Define <u>models</u>:
  - of the null hypothesis H_0: no events.
  - of the alternative hypotheses $H_1(S)$: event in region S.
- Derive a <u>score function</u>:
  - Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}$$

  - To find the most significant regions:

$$S^* = \arg\max_S F(S)$$

$$F(S) = \left(\frac{C}{B}\right)^{C} \left(\frac{C_{tot} - C}{B_{tot} - B}\right)^{C_{tot} - C} \left(\frac{C_{tot}}{B_{tot}}\right)^{-C_{tot}}$$

<u>Kulldorff's model</u>

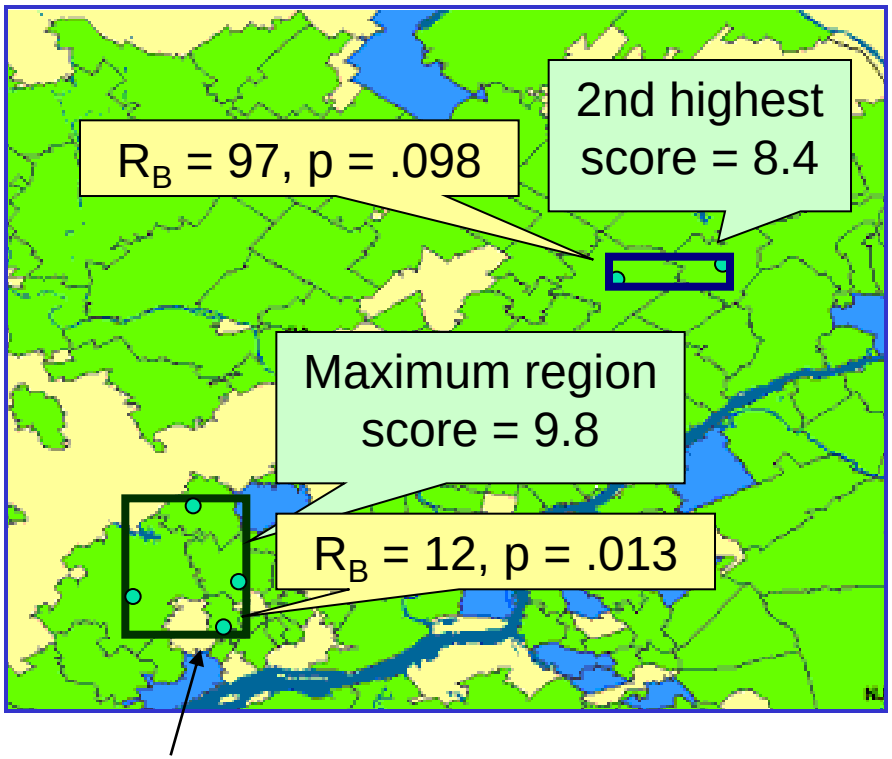
$c_i \sim \text{Poisson}(qb_i)$

H_0: $q = q_{all}$ everywhere

$H_1(S)$: $q = q_{in}$ inside S,
$q = q_{out}$ outside,
$q_{in} > q_{out}$.

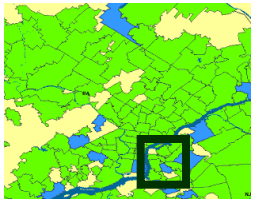
Which regions are significant?

- Randomly generate counts for R = 999 replica datasets under H_0 (i.e. assuming no events).
- Find maximum region score $F^* = \max_S F(S)$ of each replica.
- p-value of region S = $(R_B+1) / (R+1)$, where R_B = # of replicas with $F^* \geq F(S)$.
- All regions with p-values < α are significant at level α.



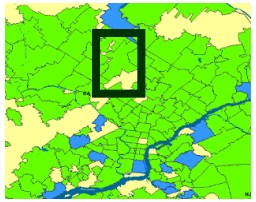
This region is significant at α = .05; no other regions are significant.

F* = 2.4



G_1

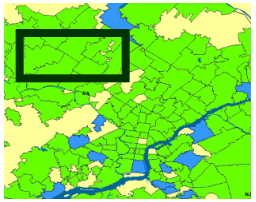
F* = 9.1



G_2

...

F* = 7.0



G_{999}

SPATIAL SCAN TIPS

1. Use historical count data, rather than population, to obtain expected counts b_i^t.

Population-based method (Kulldorff, 1997, 2001):

The old way of doing things

ED visits
OTC drug sales

Baselines represent population at risk, typically obtained from census and possibly adjusted for known risk factors.

Under the null hypothesis, we expect counts to be proportional to population.

The problem: real data doesn't behave this way!

Compare disease rate (count / pop) inside and outside region.

$q_{out} = .01$

$q_{in} = .02$

$c_i \sim Po(qb_i)$

q is disease rate, b_i is population

Different areas have different base rates

age and health of population

environmental hazards

wealth and buying habits

Base rate of an area varies over time

day of week effects holidays

seasonal trends weather

promotional sales of OTC medications

Population-based method (Kulldorff, 1997, 2001):

← The old way of doing things

ED visits
rug sales

Baselines repr
typically obt
possibly adjus

ata
way!

Under the nu
counts to be p

Compare dis
inside a

base rates
ation
ards

wealth and buying habits

<u>The solution</u>

1. Infer the time series of **expected counts** for each location, based on time series analysis of the historical data for that location.
2. Find regions where the **observed counts** are significantly higher than expected.

$q_{out} = .01$

$q_{in} = .02$

$c_i \sim Po(qb_i)$

q is disease rate, b_i is population

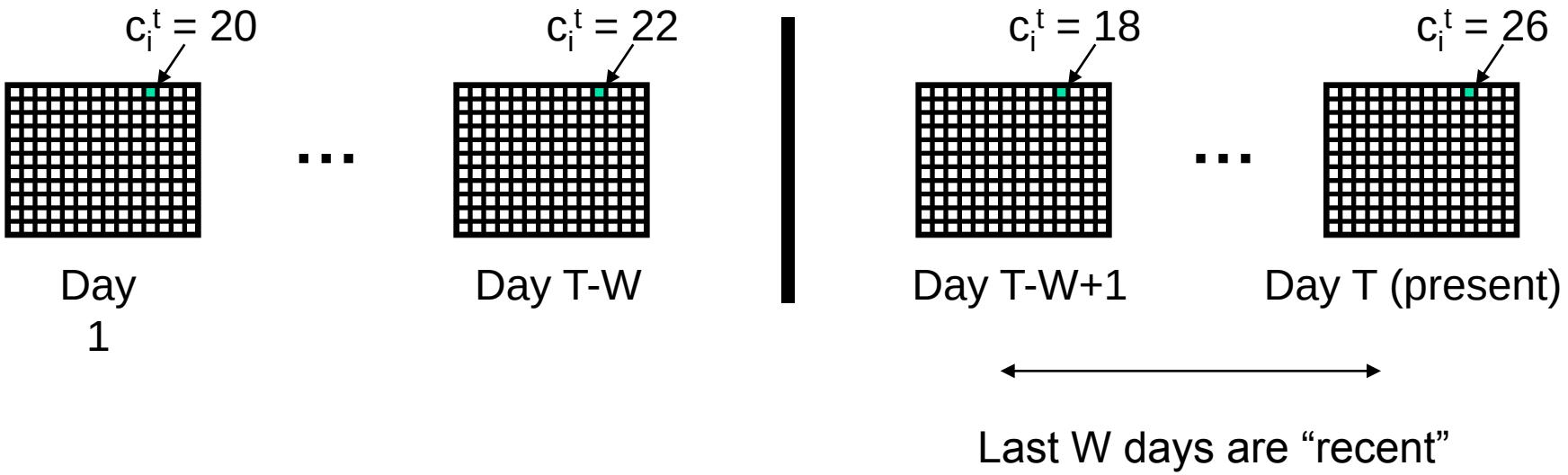
<u>Base rate of an area varies over time</u>

day of week effects — holidays

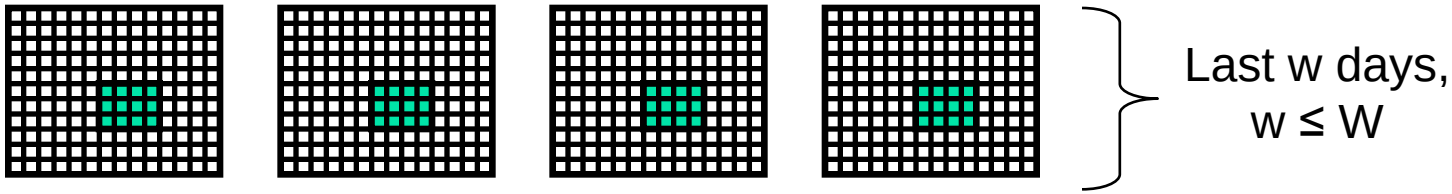
seasonal trends — weather

promotional sales of OTC medications

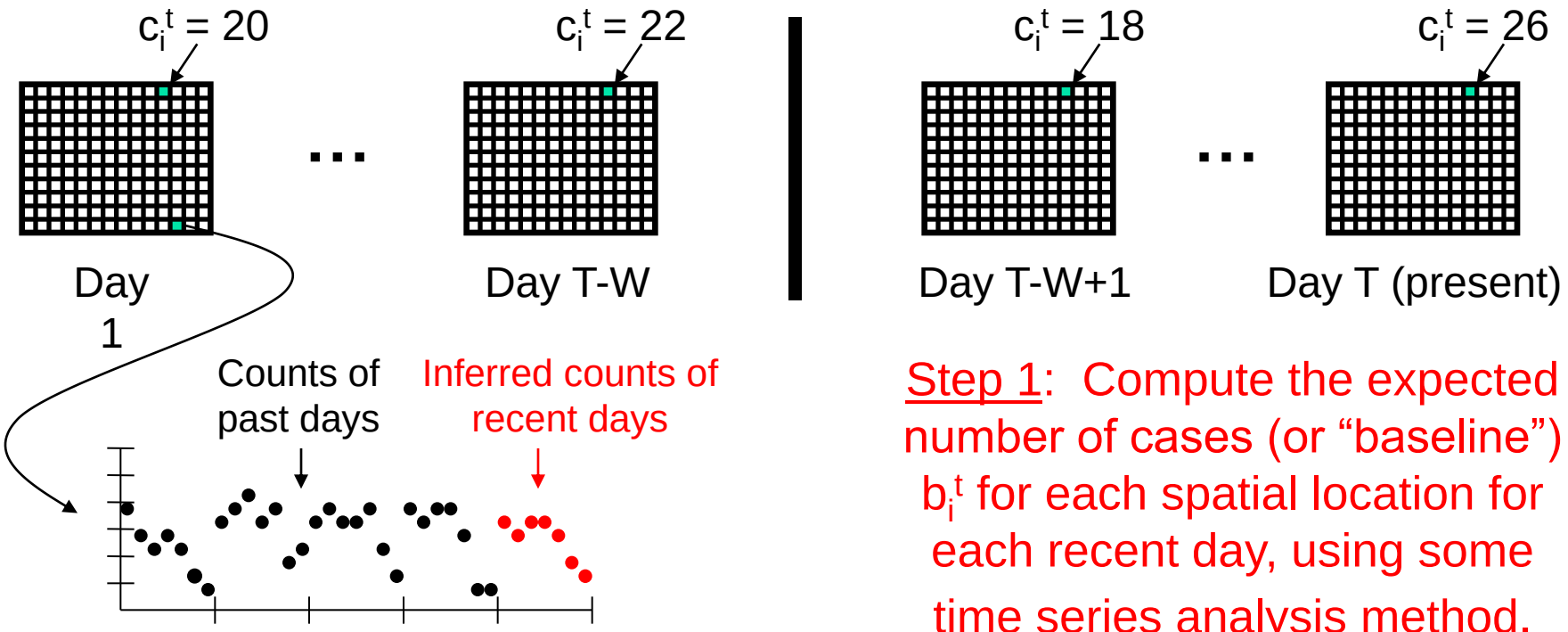
The expectation-based approach



Is there any spatial region S where the most recent counts are significantly higher than expected?



The expectation-based approach

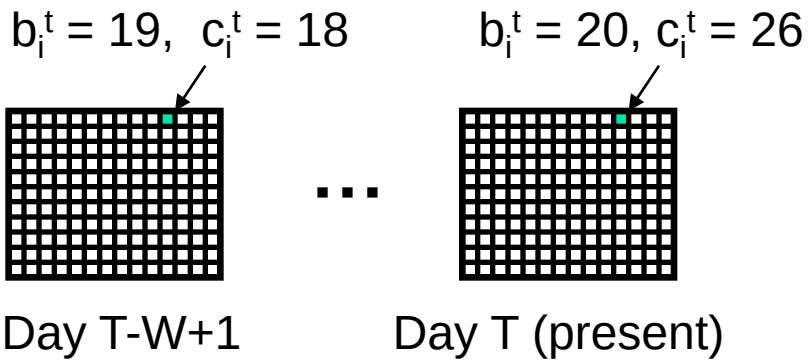


Step 1: Compute the expected number of cases (or "baseline") b_i^t for each spatial location for each recent day, using some time series analysis method.

Many other methods possible, e.g. EWLR, ARIMA, Kalman filter, Gaussian process regression.

(Weighted or unweighted) moving average. Important to adjust for **day of week** and **seasonality**.

The expectation-based approach



Step 2: use a space-time scan statistic to find clusters with the actual counts c_i^t significantly greater than the expected counts b_i^t.

To do so, we scan over the set of space-time regions $S \times \{t_{min} \ldots T\}$.

- S is a spatial region
- $t_{min} > T-W$
- Cluster ends at the present

Which variant of the scan statistic should we use?

SPATIAL SCAN TIPS

1. Use historical count data, rather than population, to obtain expected counts b_i^t.
2. Choose an appropriate likelihood ratio statistic for the given dataset and expected cluster size.

Poisson scan statistic models

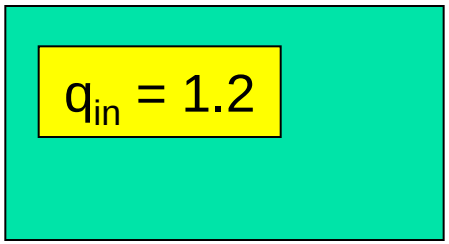
Counts are Poisson distributed: $c_i^t \sim \text{Poisson}(q_i^t b_i^t)$ —— q_i^t is relative risk, b_i^t is expected count under H_0

Expectation-based Poisson (EBP)

(Neill et al., KDD 2005)

H_0: $q_i^t = 1$ everywhere
(counts = expected)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = 1$ outside, for some $q_{in} > 1$.
(counts > expected in S)

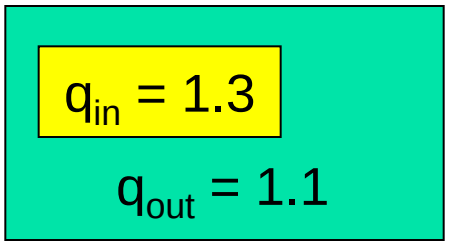


Kulldorff's scan statistic (KULL)

(Kulldorff, 1997, 2001)

H_0: $q_i^t = q_{all}$ everywhere
(inside = outside)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = q_{out}$ outside, for some $q_{in} > q_{out}$.
(inside > outside)



Poisson scan statistic models

Counts are Poisson distributed: $c_i^t \sim \text{Poisson}(q_i^t b_i^t)$ —— q_i^t is relative risk, b_i^t is expected count under H_0

Expectation-based Poisson (EBP)

(Neill et al., KDD 2005)

H_0: $q_i^t = 1$ everywhere
(counts = expected)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = 1$ outside, for some $q_{in} > 1$.
(counts > expected in S)

$$F(S) = \left(\frac{C}{B}\right)^C e^{B-C}$$

(if C > B)

Kulldorff's scan statistic (KULL)

(Kulldorff, 1997, 2001)

H_0: $q_i^t = q_{all}$ everywhere
(inside = outside)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = q_{out}$ outside, for some $q_{in} > q_{out}$.
(inside > outside)

$$F(S) = \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}} \left(\frac{C_{all}}{B_{all}}\right)^{-C_{all}}$$

(if $C_{in} / B_{in} > C_{out} / B_{out}$)

Gaussian scan statistic models

Counts are Gaussian distributed: $c_i^t \sim \text{Gaussian}(q_i^t b_i^t, \sigma_i^t)$

Let $C' = \Sigma\, c_i^t b_i^t / (\sigma_i^t)^2$ and $B' = \Sigma\, (b_i^t)^2 / (\sigma_i^t)^2$

Expectation-based Gaussian (EBG)

(Neill, Ph.D. thesis, 2006)

H_0: $q_i^t = 1$ everywhere
(counts = expected)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = 1$ outside, for some $q_{in} > 1$.
(counts > expected in S)

$$F(S) = \exp\left(\frac{(C')^2}{2B'} + \frac{B'}{2} - C'\right)$$

(if $C' > B'$)

Population-based Gaussian (PBG)

(Neill, Ph.D. thesis, 2006)

H_0: $q_i^t = q_{all}$ everywhere
(inside = outside)

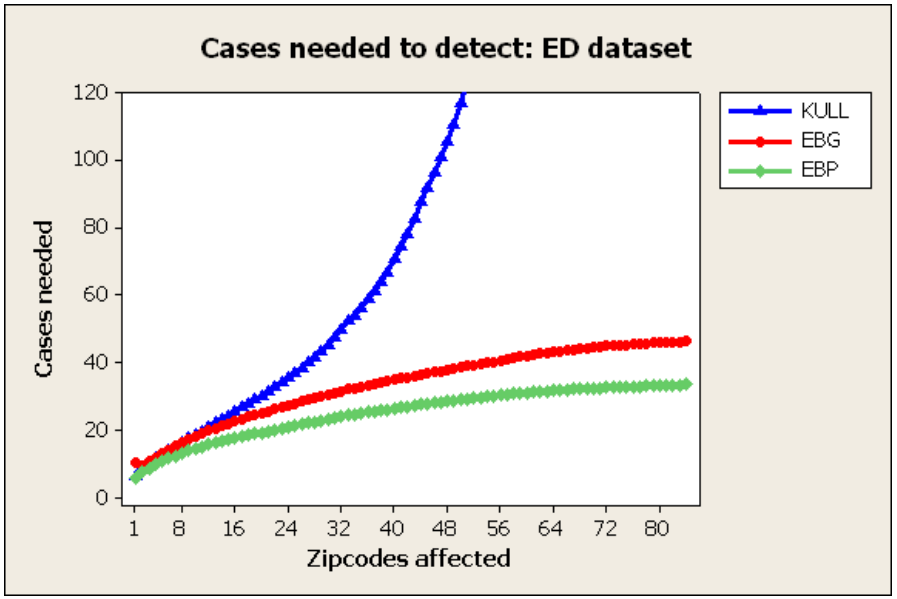
$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = q_{out}$ outside, for some $q_{in} > q_{out}$.
(inside > outside)

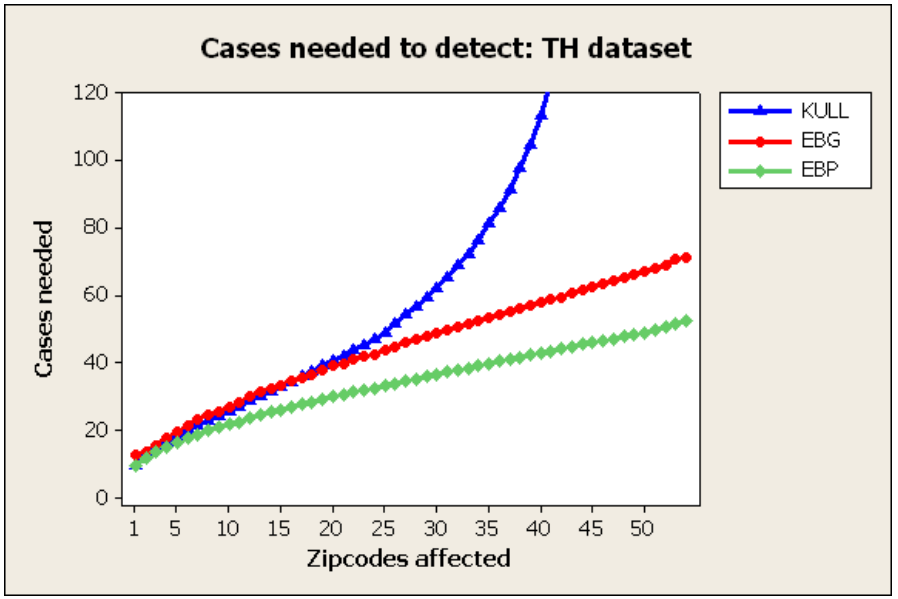
$$F(S) = \exp\left(\frac{C'_{in}{}^2}{2B'_{in}} + \frac{C'_{out}{}^2}{2B'_{out}} - \frac{C'_{all}{}^2}{2B'_{all}}\right)$$

(if $C'_{in} / B'_{in} > C'_{out} / B'_{out}$)

Comparison of detection power

We computed the average number of injected cases needed for each method to detect 90% of outbreaks on a given day, as a function of the number of affected zip codes.

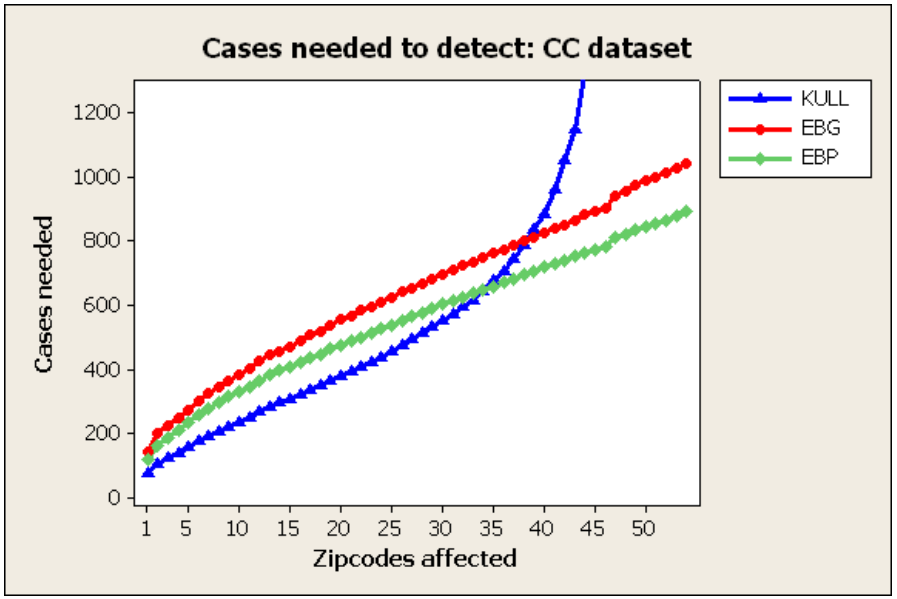


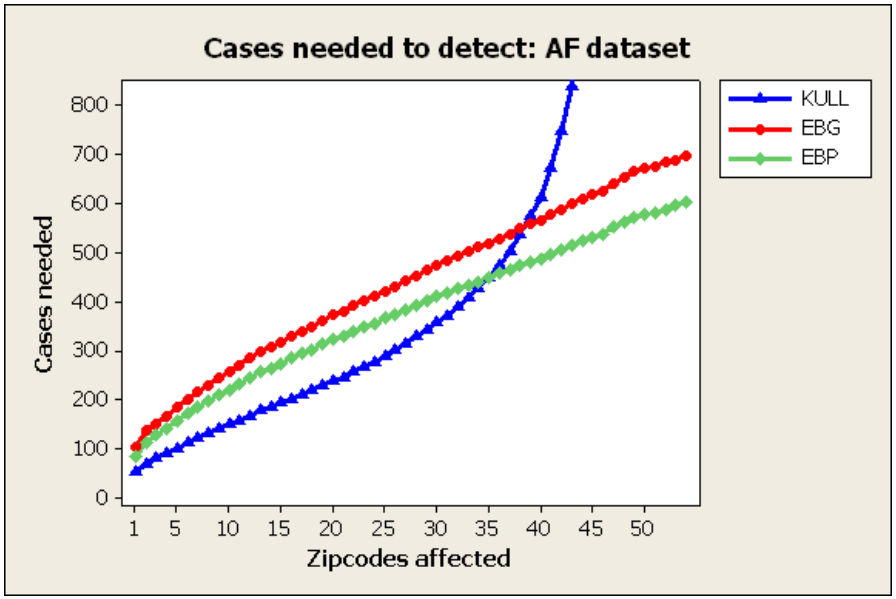


Respiratory ED visits and thermometer sales: EBP achieves consistently high performance. KULL has low detection power for large outbreaks.

Comparison of detection power

We computed the average number of injected cases needed for each method to detect 90% of outbreaks on a given day, as a function of the number of affected zip codes.

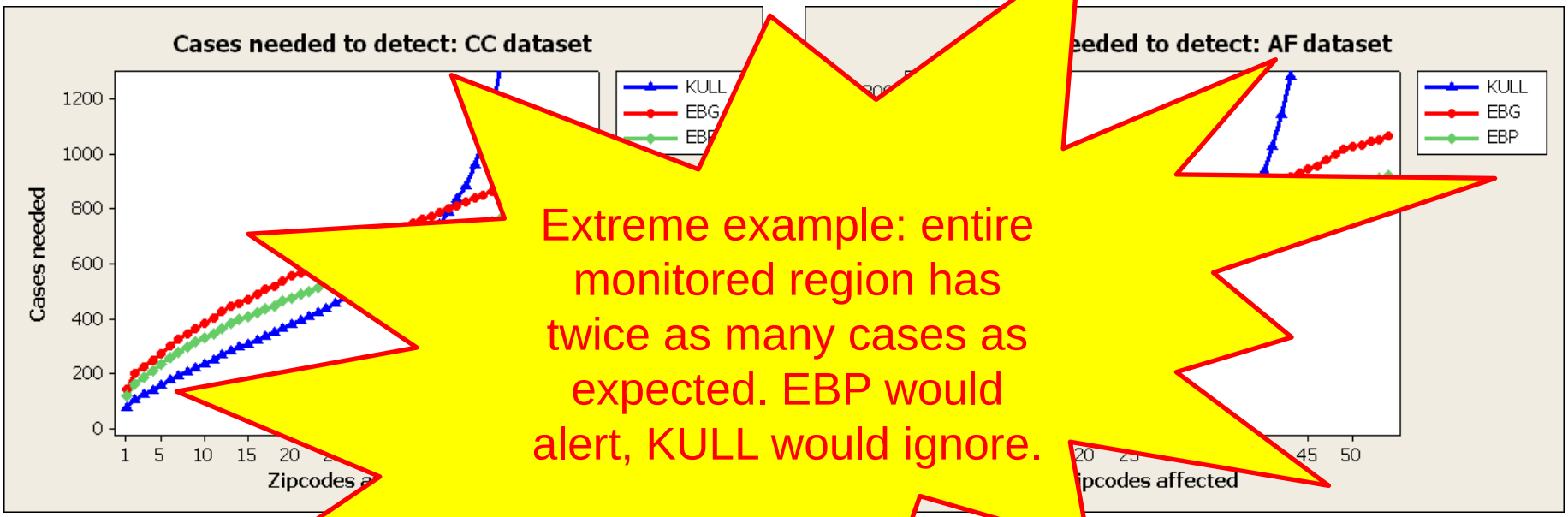




Cough/cold and anti-fever OTC sales: KULL outperforms EBP when less than 2/3 of zip codes are affected, but has low power for large outbreaks.

Comparison of detection power

We computed the average number of injected cases needed for each method to detect 90% of outbreaks on a given day, as a function of the number of affected zip codes.



Cough/cold and anti- OTC sales: KULL outperforms EBP when less than 2/3 of zip codes are affected, but has low power for large outbreaks.

SPATIAL SCAN TIPS

1. Use historical count data, rather than population, to obtain expected counts b_i^t.
2. Choose an appropriate likelihood ratio statistic for the given dataset and expected cluster size.
3. Decide whether randomization testing is a good idea (typically the answer is NO!)

Randomization testing considerations

- Randomization is one way to provide a suggested **threshold** for sounding the alarm, but there are other options as well.
  - Each day, report top-k non-overlapping clusters.
  - Report all clusters with scores over fixed value.

  → Use the **empirical distribution** of maximum scores from historical data (i.e. to be significant at α = .05, must beat ~95% of historical days).
- Randomization multiplies **computation time** by the number of Monte Carlo replications (typically at least 100, often 1000 or 10,000).

Randomization testing considerations

- Randomization testing identifies clusters which are unexpected given the null hypothesis… but H_0 makes many assumptions we don't believe.
  - Independent Poisson-distributed counts (not overdispersed, no spatial autocorrelation, etc.)
  - No irrelevant anomalies (data entry errors, etc.)
  - Uniform risk assumed under the null: baselines capture all the variation in counts if no outbreaks are occurring.
- Randomization guarantees the desired FPR (e.g. $\alpha = .05$) if the null is true, but not otherwise.
- In real data (with incorrectly specified null) FPR is **much** higher: 11-57% at $\alpha = .05$ for OTC data.

Randomization testing considerations

- High false positive rate can harm detection power (days to detect for a given FPR, e.g. 1/month).
  - Many days with p-values of 1/(R+1) → indistinguishable.
  - On the ED and OTC datasets, reporting the regions with lowest p-values gave much **lower** detection power than simply reporting the highest-scoring regions.
  - Using the Gumbel p-value correction helps, but still does not achieve higher detection power.
- Randomization often does not help, and can even harm, performance. So when can it be helpful?
  - Insufficient historical data to use empirical scores.
  - Major changes in empirical score distribution over time due to population shifts, new monitored locations, etc.

SPATIAL SCAN TIPS

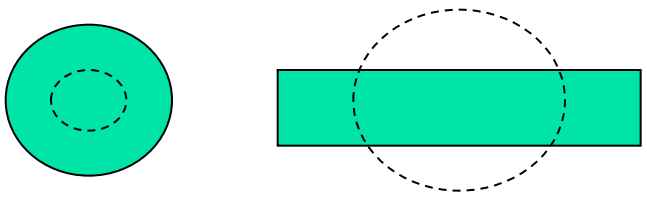
1. Use historical count data, rather than population, to obtain expected counts b_i^t.
2. Choose an appropriate likelihood ratio statistic for the given dataset and expected cluster size.
3. Decide whether randomization testing is a good idea (typically the answer is NO!)
4. Choose appropriate set of search regions.

Choosing the set of search regions

- Some practical considerations:
  - Set of regions should cover entire search space.
  - Regions should overlap, not partition the space.
- Choose a set of regions that corresponds well with the size/shape of the clusters we want to detect.
  - Typical approaches consider some fixed shape (circles, rectangles) and vary the location and dimensions.

Don't search too few regions:

Reduced power to detect clusters outside the search space.



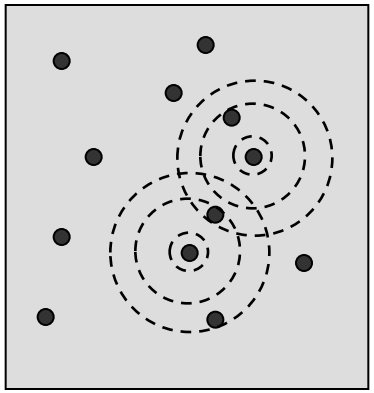
Don't search too many regions:

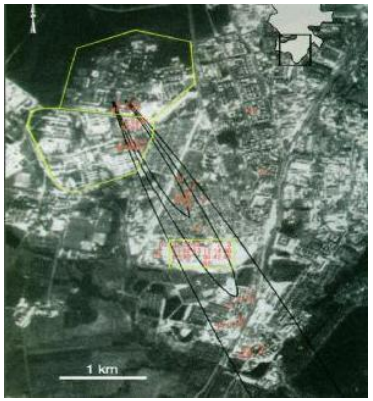
Overall power to detect any given subset of regions reduced because of multiple hypothesis testing.

Computational infeasibility!

Choosing the set of search regions

- Kulldorff's original spatial scan searches over circular regions of varying radius, centered at each spatial location s_i.
- Since the score function F(S) depends only on which locations are included, we need to search $O(N^2)$ regions, each consisting of a center location and its k-NN.
- Advantages: computationally efficient, generalizable to arbitrary metric spaces, high detection power for compact clusters.
- Disadvantage: low power for elongated/irregular clusters.





April 1979: inadvertent release of anthrax from a Soviet biological weapons facility, 77 cases confirmed.

Disease cluster elongated due to wind.

Choosing the set of search regions

- Kulldorff's original spatial scan searches over circular regions of varying radius, centered at each spatial location s_i.
- Since the score function F(S) depends only on which locations are included, we need to search $O(N^2)$ regions, each consisting of a center location and its k-NN.
- Advantages: computationally efficient, generalizable to arbitrary metric spaces, high detection power for compact clusters.
- Disadvantage: low power for elongated/irregular clusters.

Many recent spatial scan variants search over elongated clusters, e.g. rectangles[1] or ellipses[2]

Other variants: heuristic search over all connected regions[3], or exhaustive search over a subset of connected regions[4,5]

Main challenge: efficient computation!

[1]Neill and Moore, KDD 2004

[2]Kulldorff et al., Stat. Med., 2007

[3]Duczmal and Assuncao, CSDA, 2004

[4]Tango and Takahashi, IJHG, 2005

[5]Patil and Taillie, EES, 2004

Choosing the set of search regions

- Kulldorff's original spatial scan searches over circular regions of varying radius, centered at each spatial location s_i.
- Since the score function F(S) depends only on which locations are included, we need to search $O(N^2)$ regions, each consisting of a center location and its k-NN.

Many recent spatial scan variants search over elongated clusters, e.g. rectangles[1] or ellipses[2]

Other variants: heuristic search over all connected regions[3], or exhaustive search over a subset of connected regions[4,5]

Our recently proposed "Linear-Time Subset Scanning" methods enable efficient optimization over irregularly shaped clusters, finding the highest-scoring proximity-constrained subsets of locations, and substantially improving detection time and accuracy.

Main challenge: efficient computation!

[1]Neill and Moore, KDD 2004

[2]Kulldorff et al., Stat. Med., 2007

[3]Duczmal and Assuncao, CSDA, 2004

[4]Tango and Takahashi, IJHG, 2005

[5]Patil and Taillie, EES, 2004

SPATIAL SCAN TIPS

1. Use historical count data, rather than population, to obtain expected counts b_i^t.
2. Choose an appropriate likelihood ratio statistic for the given dataset and expected cluster size.
3. Decide whether randomization testing is a good idea (typically the answer is NO!)
4. Choose appropriate set of search regions.
5. Simpler is not always better… see what the recent literature has to offer.

Multivariate models

- Timeliness and accuracy of detection can often be dramatically improved by combining information from **multiple data streams**.
  - Lots of recent work here- parametric, nonparametric, Bayesian…
- Multivariate Bayesian Scan Statistic
  - MBSS allows us to model and differentiate between **multiple outbreak types**, as well as distinguishing between outbreaks and false positive alerts (e.g. promotional OTC sales) → event characterization.
- Future advances will continue to improve the timeliness, accuracy, and scalability of spatial event detection methods.

References

- L. Duczmal and R. Assuncao. A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis*, 45:269–286, 2004.
- M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6): 1481–1496, 1997.
- M. Kulldorff. Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society A*, 164: 61–72, 2001.
- M. Kulldorff, L. Huang, L. Pickle, and L. Duczmal. An elliptic spatial scan statistic. *Statistics in Medicine*, 25:3929–3943, 2006.
- M. Kulldorff, F. Mostashari, L. Duczmal, W. K. Yih, K. Kleinman, and R. Platt. Multivariate scan statistics for disease surveillance. *Statistics in Medicine*, 26: 1824–1833, 2007.
- D.B. Neill, A.W. Moore, M. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. *Proc. 11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 218-227, 2005.
- D.B. Neill. Detection of spatial and spatio-temporal clusters. Ph.D. thesis, Carnegie Mellon University, Department of Computer Science, 2006.
- D.B. Neill and J. Lingwall. A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance* 4: 106, 2007.
- D.B. Neill. Fast and flexible outbreak detection by linear-time subset scanning. *Advances in Disease Surveillance* 5: 48, 2008.
- **D.B. Neill. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics* 8: 20, 2009.**
- D.B. Neill. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, 2009, 25: 498-517.
- D.B. Neill and G.F. Cooper. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning*, 2010, 79: 261-282.
- G. P. Patil and C. Taillie. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Envir. Ecol. Stat*., 11: 183–197, 2004.
- T. Tango and K. Takahashi. A flexibly shaped spatial scan statistic for detecting clusters. *Intl. Journal of Health Geographics*, 4: 11, 2005.

Thanks!!! Questions???

1. Use historical count data, rather than population, to obtain expected counts b_i^t.
2. Choose an appropriate likelihood ratio statistic for the given dataset and expected cluster size.
3. Decide whether randomization testing is a good idea (typically the answer is NO!)
4. Choose appropriate set of search regions.
5. Simpler is not always better… see what the recent literature has to offer.