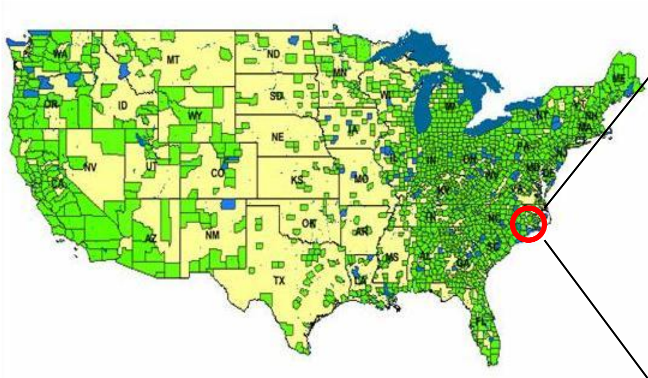
Generalized Fast Subset Sums for Bayesian Detection and Visualization

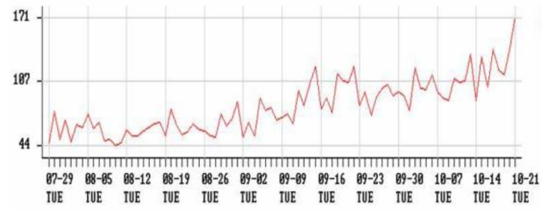
Daniel B. Neill* and Yandong Liu
Event and Pattern Detection Laboratory
Carnegie Mellon University
{neill, yandongl} @ cs.cmu.edu

This work was partially supported by NSF grants
IIS-0916345, IIS-0911032, and IIS-0953330.

Multivariate event detection



Daily health data from thousands of hospitals and pharmacies nationwide.



Time series of counts $c_{i,m}^t$ for each zip code s_i for each data stream d_m.

d_1 = respiratory ED

d_2 = constitutional ED

d_3 = OTC cough/cold

d_4 = OTC anti-fever

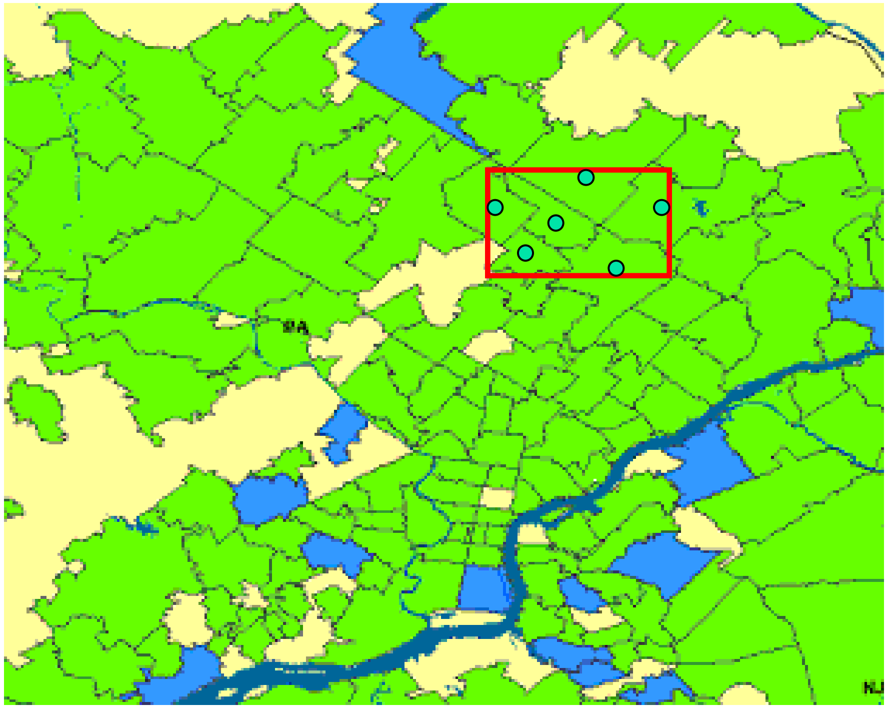
etc.

Given all of this nationwide health data on a daily basis, we want to obtain a complete situational awareness by integrating information from the multiple data streams.

More precisely, we have three main goals: to detect any emerging events (i.e. outbreaks of disease), characterize the type of event, and pinpoint the affected areas.

Expectation-based scan statistics

(Kulldorff, 1997; Neill and Moore, 2005)

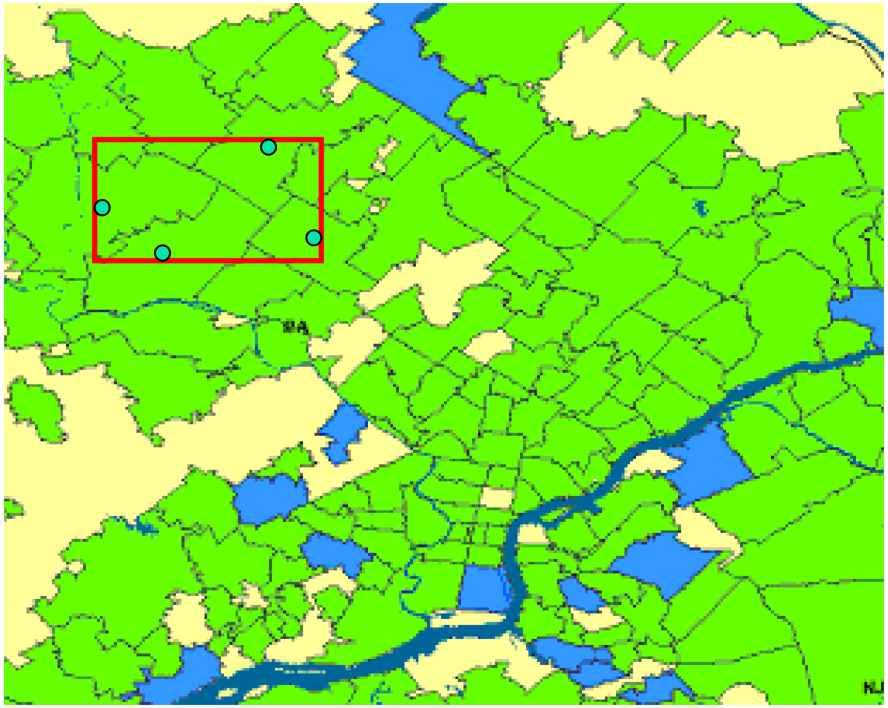


To detect and localize events, we can search for space-time regions where the number of cases is higher than expected.

Imagine moving a window around the scan area, allowing the window size, shape, and temporal duration to vary.

Expectation-based scan statistics

(Kulldorff, 1997; Neill and Moore, 2005)

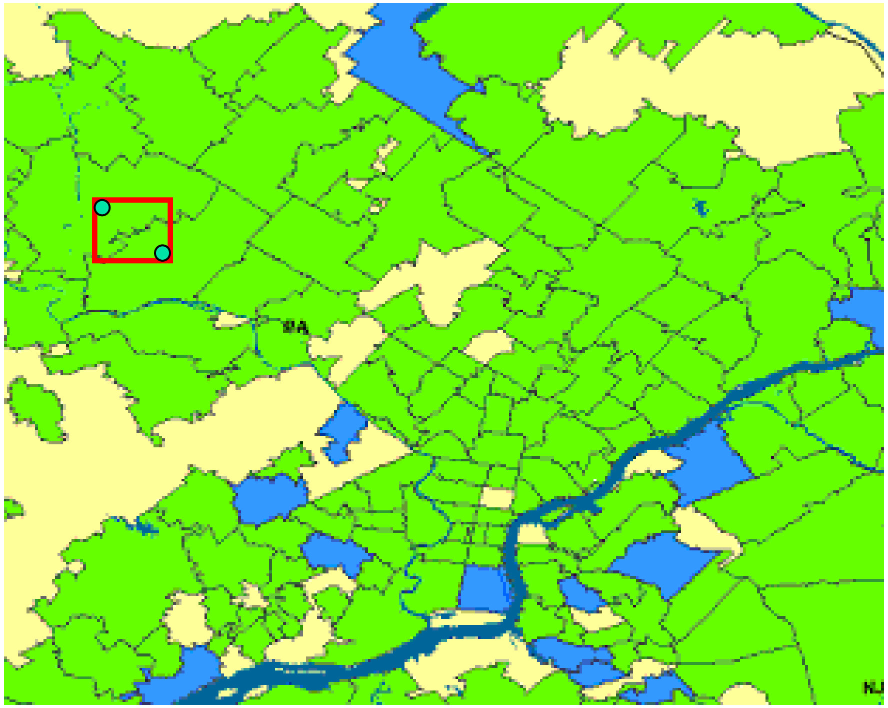


To detect and localize events, we can search for space-time regions where the number of cases is higher than expected.

Imagine moving a window around the scan area, allowing the window size, shape, and temporal duration to vary.

Expectation-based scan statistics

(Kulldorff, 1997; Neill and Moore, 2005)

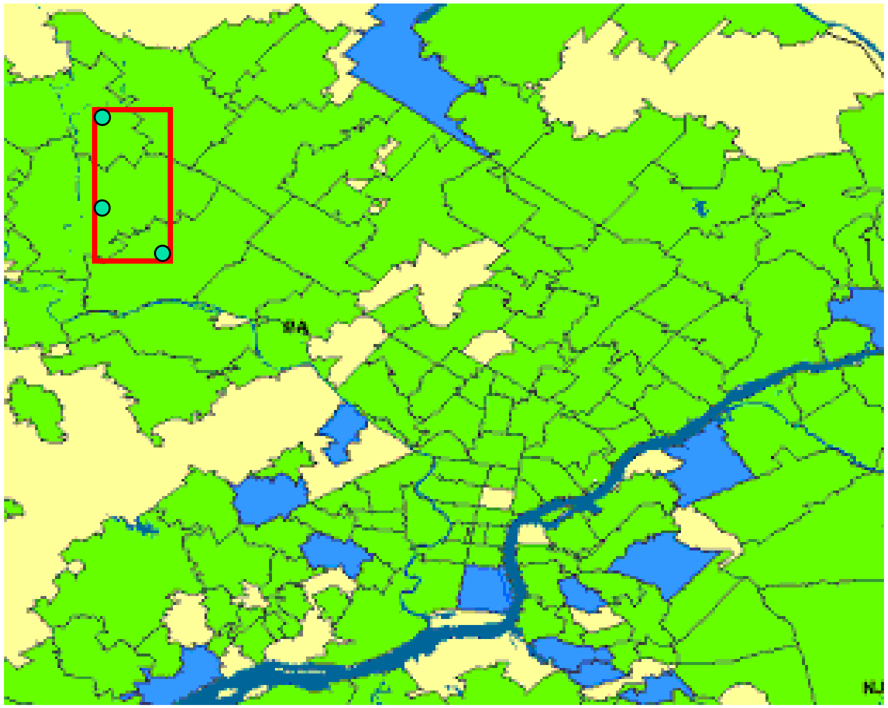


To detect and localize events, we can search for space-time regions where the number of cases is higher than expected.

Imagine moving a window around the scan area, allowing the window size, shape, and temporal duration to vary.

Expectation-based scan statistics

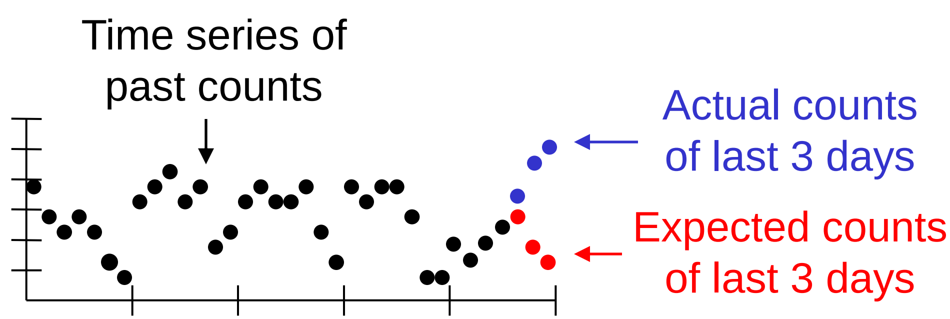
(Kulldorff, 1997; Neill and Moore, 2005)



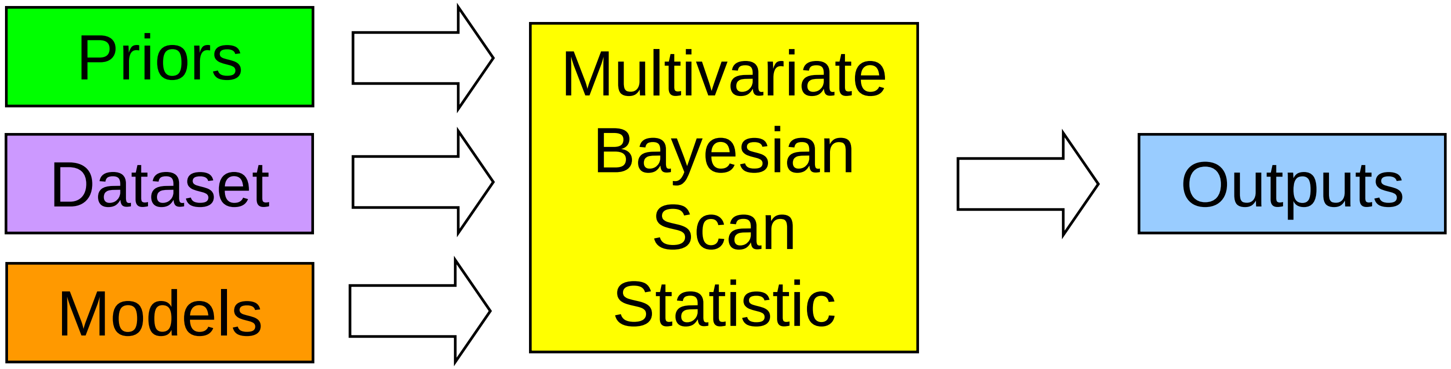
To detect and localize events, we can search for <u>space-time regions</u> where the number of cases is higher than expected.

Imagine moving a window around the scan area, allowing the window size, shape, and temporal duration to vary.

For each subset of locations, we examine the aggregated time series, and compare actual to expected counts.



Overview of the MBSS method

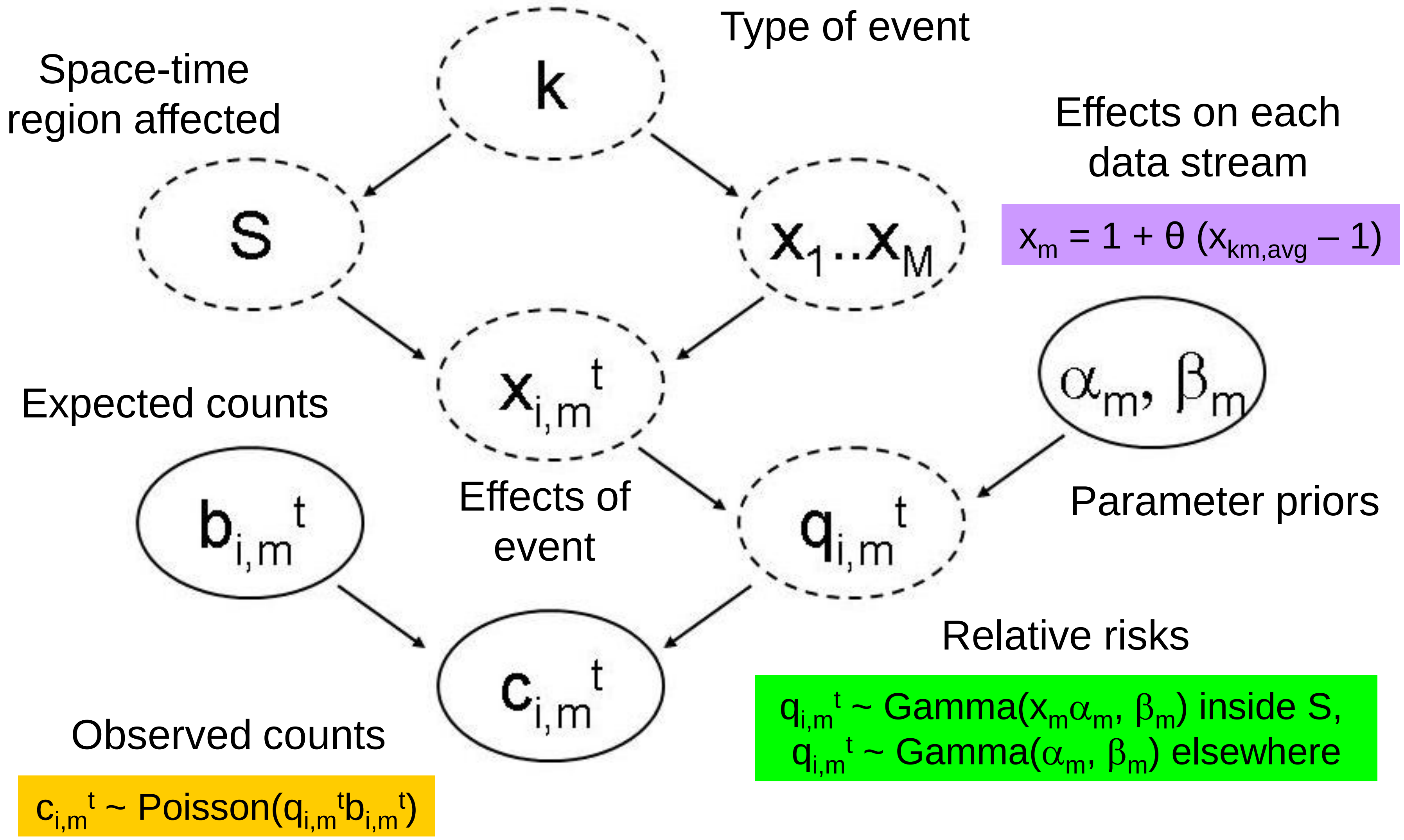


Given a set of event types E_k, a set of space-time regions S, and the multivariate dataset D, MBSS outputs the posterior probability $Pr(H_1(S, E_k) \mid D)$ of each type of event in each region, as well as the probability of no event, $Pr(H_0 \mid D)$.

We must provide the prior probability $Pr(H_1(S, E_k))$ of each event type E_k in each region S, as well as the prior probability of no event, $Pr(H_0)$.

MBSS uses Bayes' Theorem to combine the data likelihood given each hypothesis with the prior probability of that hypothesis: $Pr(H \mid D) = Pr(D \mid H)\, Pr(H) / Pr(D)$.

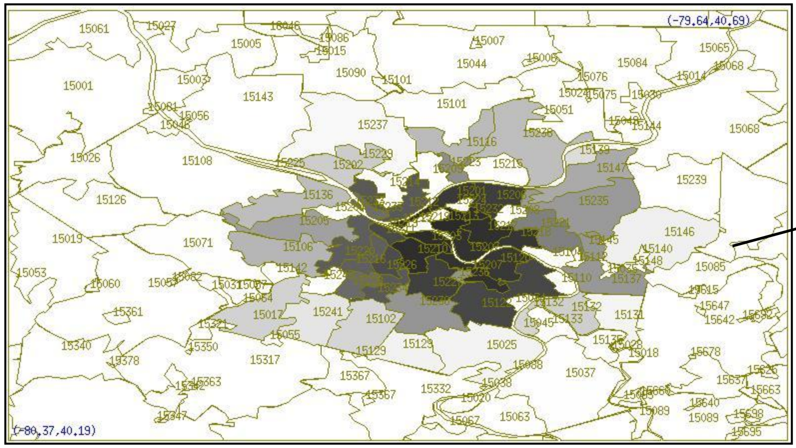
The Bayesian hierarchical model



Interpretation and visualization

MBSS gives the total posterior probability of each event type E_k, and the distribution of this probability over space-time regions S.

<u>Visualization</u>: $Pr(H_1(s_i, E_k)) = \sum Pr(H_1(S, E_k))$ for all regions S containing location s_i.



<u>Posterior probability map</u>

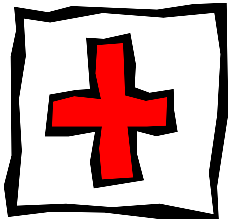
Total posterior probability of a respiratory outbreak in each Allegheny County zip code.

Darker shading = higher probability.

MBSS: advantages and limitations

MBSS can detect faster and more accurately by integrating multiple data streams.

MBSS can model and differentiate between multiple potential causes of an event.



MBSS assumes a uniform prior for circular regions and zero prior for non-circular regions, resulting in low power for **elongated** or **irregular** clusters.

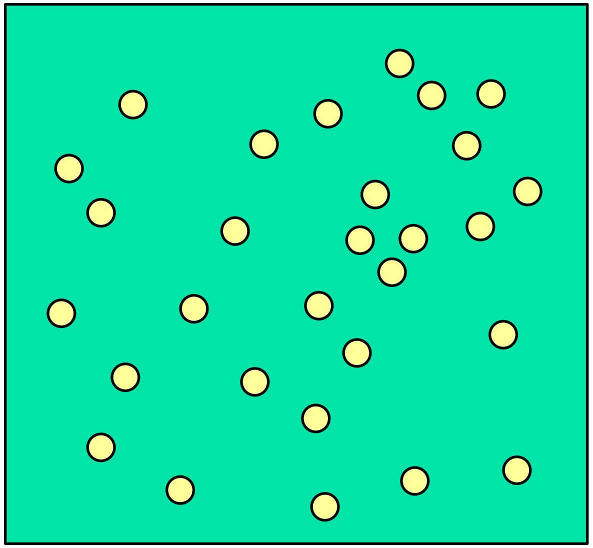
There are too many subsets of the data (2^N) to compute likelihoods for all of them!

How can we extend MBSS to **efficiently** detect irregular clusters?

Generalized Fast Subset Sums

We define a non-uniform prior $\Pr(H_1(S, E_k))$ over all 2^N subsets of the data.

This prior has hierarchical structure:

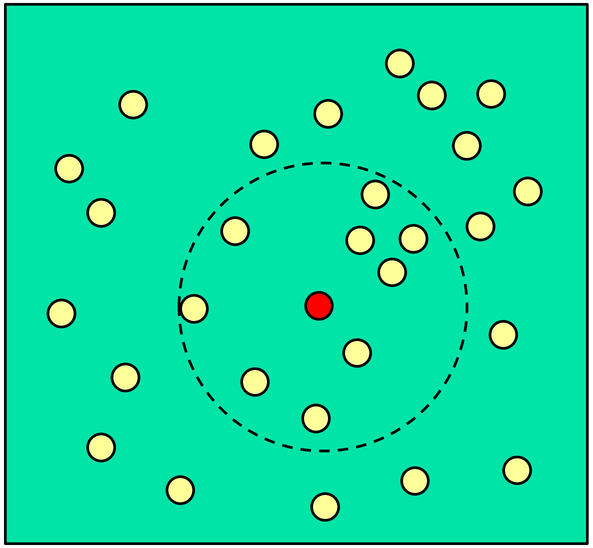


Generalized Fast Subset Sums

We define a non-uniform prior $Pr(H_1(S, E_k))$ over all 2^N subsets of the data.

This prior has hierarchical structure:

1. Choose **center location s_c** from $\{s_1 \ldots s_N\}$, given multinomial $Pr(s_i)$.
2. Choose **neighborhood size n** from $\{1 \ldots n_{max}\}$, given multinomial $Pr(n)$.

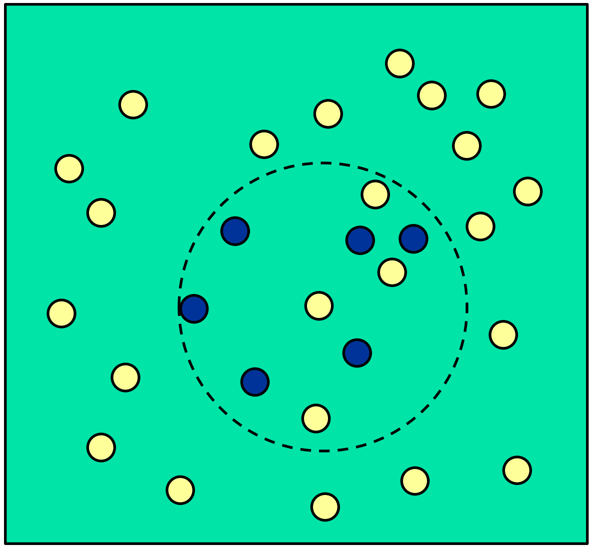


Generalized Fast Subset Sums

We define a non-uniform prior $Pr(H_1(S, E_k))$ over all 2^N subsets of the data.

This prior has hierarchical structure:

1. Choose **center location s_c** from $\{s_1 \ldots s_N\}$, given multinomial $Pr(s_i)$.
2. Choose **neighborhood size n** from $\{1 \ldots n_{max}\}$, given multinomial $Pr(n)$.
3. For each $s_i \in S_{cn}$, include s_i in S with probability p, for a fixed $0 < p \leq 1$.



This prior distribution has non-zero prior probabilities for any given subset S, but more compact clusters have larger priors.

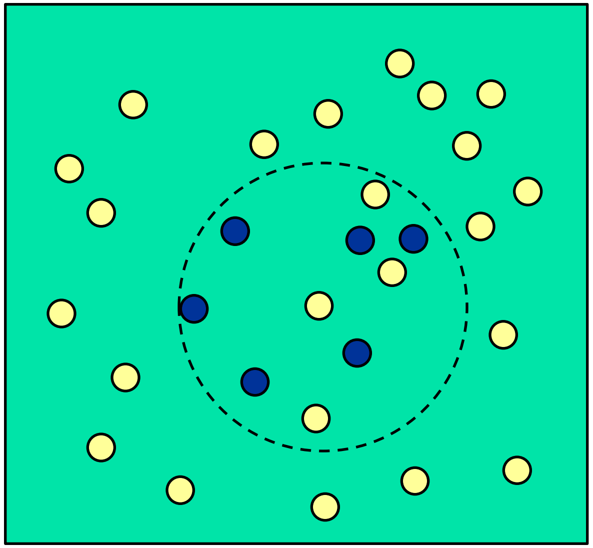
Parameter p controls the sparsity of detected clusters.
Large p = compact clusters. Small p = dispersed clusters.

Generalized Fast Subset Sums

We define a non-uniform prior $Pr(H_1(S, E_k))$ over all 2^N subsets of the data.

This prior has hierarchical structure:

1. Choose **center location s_c** from $\{s_1 \ldots s_N\}$, given multinomial $Pr(s_i)$.
2. Choose **neighborhood size n** from $\{1 \ldots n_{max}\}$, given multinomial $Pr(n)$.
3. For each $s_i \in S_{cn}$, include s_i in S with probability p, for a fixed $0 < p \leq 1$.



p = 0.5 corresponds to the original Fast Subset Sums approach described in (Neill, 2010), assuming that all subsets are equally likely given the neighborhood.

p = 1 corresponds to MBSS, searching circular regions only.

Generalized **Fast** Subset Sums

Naïve computation of posterior probabilities using this prior requires summing over an exponential number of regions, which is infeasible.

However, the total posterior probability of an outbreak, $Pr(H_1(E_k) \mid D)$, and the posterior probability map, $Pr(H_1(s_i, E_k) \mid D)$, can be calculated efficiently **without** computing the probability of each region S.

In the original MBSS method, the **likelihood ratio** of spatial region S for a given event type E_k and event severity θ can be found by multiplying the likelihood ratios $LR(s_i \mid E_k, \theta)$ for all locations s_i in S.

In GFSS, the **average likelihood ratio** of the 2^n subsets for a given center s_c and neighborhood size n can be found by multiplying the quantities $(p \times LR(s_i \mid E_k, \theta) + (1-p))$ for all locations s_i in S.

Since the prior is uniform for a given center and neighborhood, we can compute the posteriors for each s_c and n, and marginalize over them.

Evaluation

- We injected simulated disease outbreaks into two streams of Emergency Department data (cough, nausea) from 97 Allegheny County zip codes.
- Results were computed for ten different outbreak shapes, including compact, elongated, and irregularly-shaped, with 200 injects of each type.
- We evaluated GFSS (with varying p) in terms of run time, timeliness of detection, proportion of outbreaks detected, and spatial accuracy.

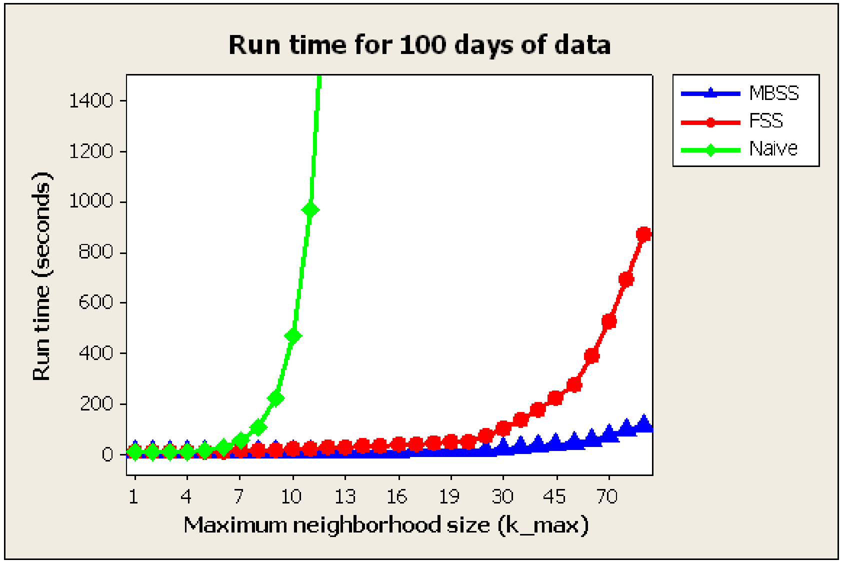
Computation time

We compared the run times of MBSS, GFSS, and a naïve subset sums implementation as a function of the maximum neighborhood size n_{max}.

Run time of MBSS increased gradually with increasing n_{max}, up to 1.2 seconds per day of data.

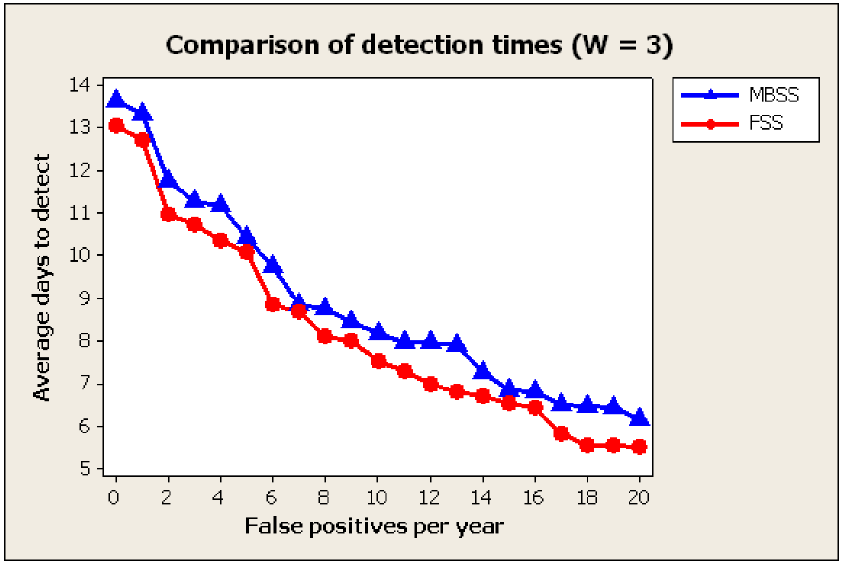
Run time of Naïve Subset Sums increased exponentially, making it infeasible for $n_{max} \geq 25$.

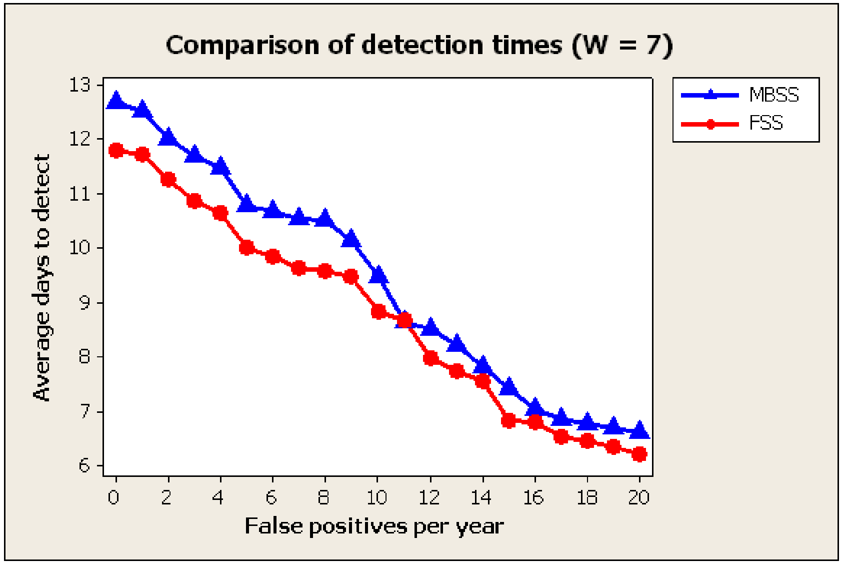
Run time of GFSS scaled quadratically with n_{max}, up to 8.8 seconds per day of data.



Thus, while GFSS is approximately 7.5x slower than the original MBSS method, it is still extremely fast, computing the posterior probability map for each day of data in under nine seconds.

Timeliness of detection

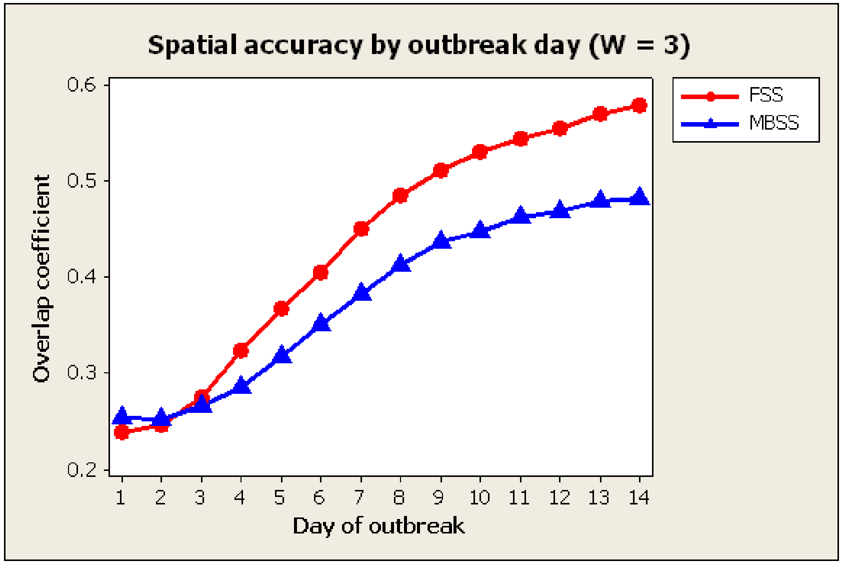


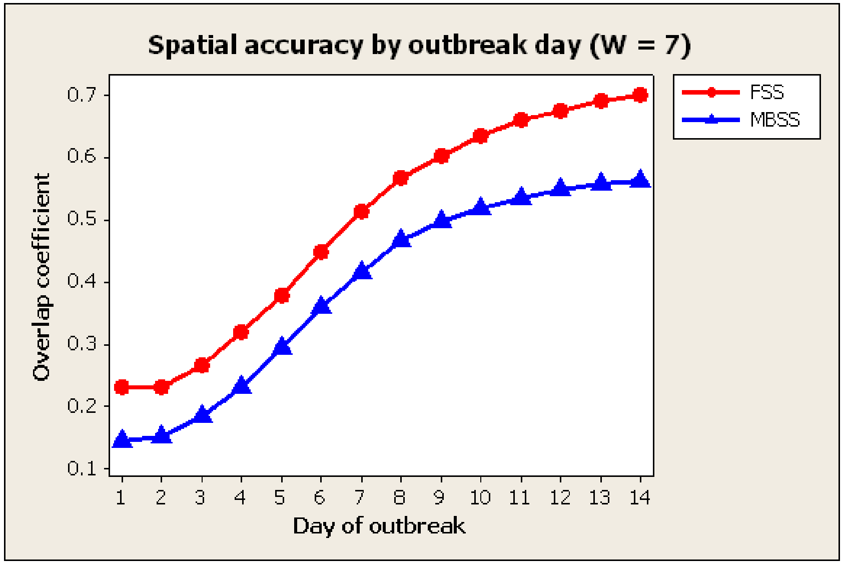


With p = 0.5, GFSS detected an average of **one day earlier** than MBSS for maximum temporal window W = 3, and **0.54 days earlier** for W = 7, with less than half as many missed outbreaks.

Both methods achieve similar detection times for compact outbreak regions. For highly elongated outbreaks, GFSS detects 1.3 to 2.2 days earlier, and for irregular regions, GFSS detects 0.3 to 1.2 days earlier.

Spatial accuracy





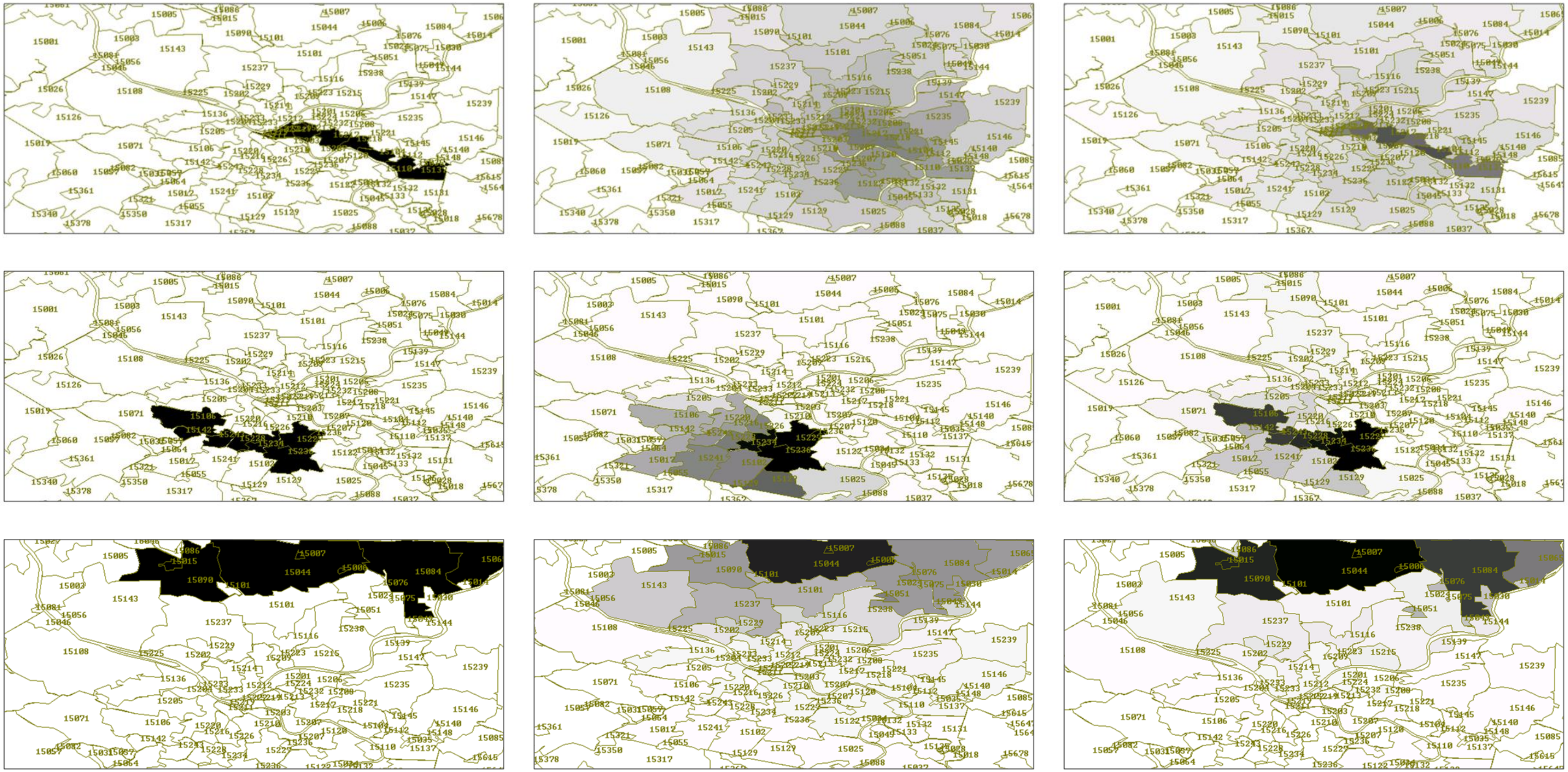
As measured by the overlap coefficient between true and detected clusters, GFSS outperformed MBSS by 10-15%.

For elongated and irregular clusters, GFSS had much higher precision and recall. For compact clusters, GFSS had higher precision, and MBSS had higher recall.

Posterior probability maps

GFSS has much higher spatial accuracy than MBSS for elongated clusters.

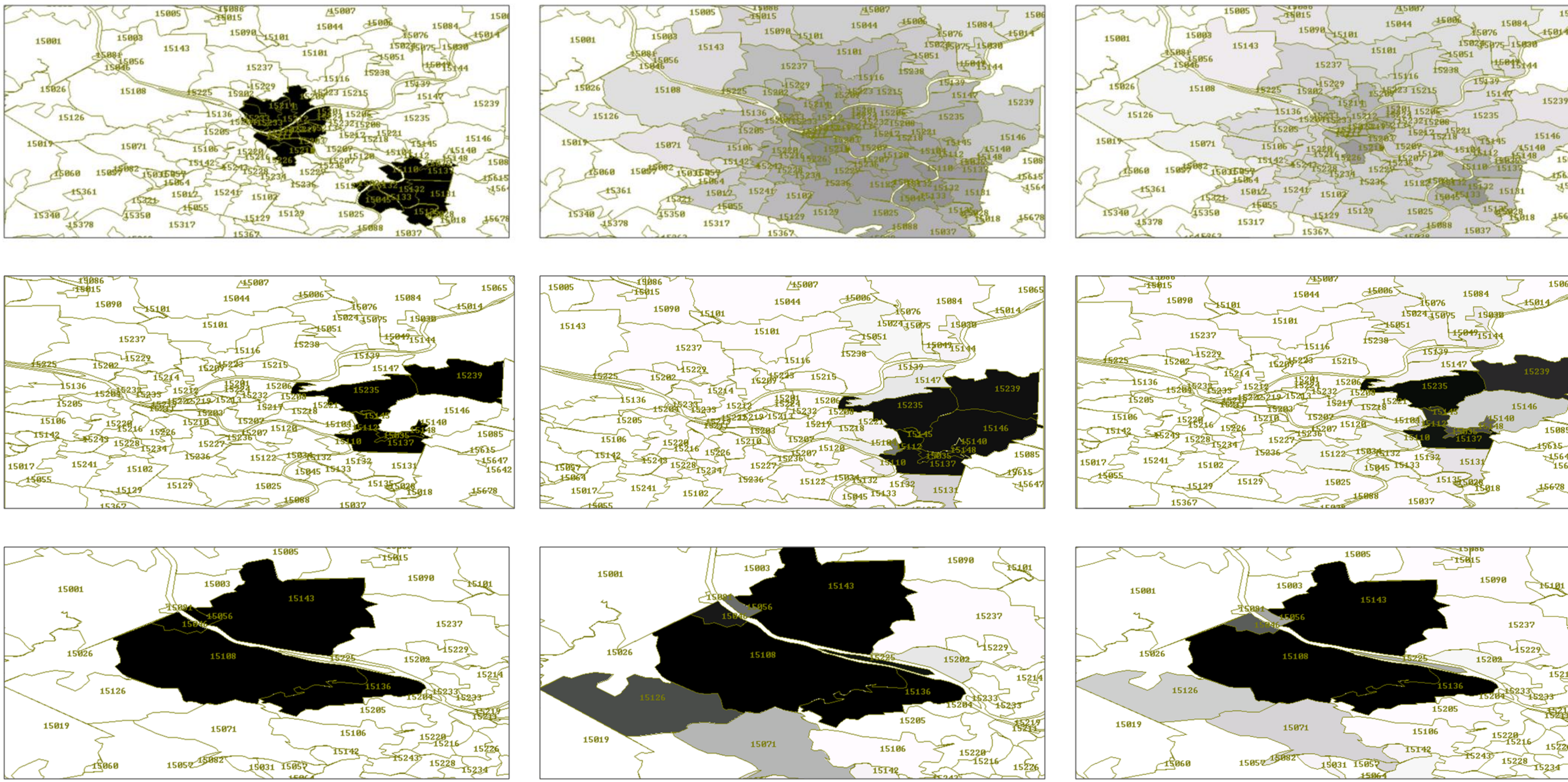
True outbreak region | MBSS (p = 1) | GFSS (p = 0.5)



Posterior probability maps

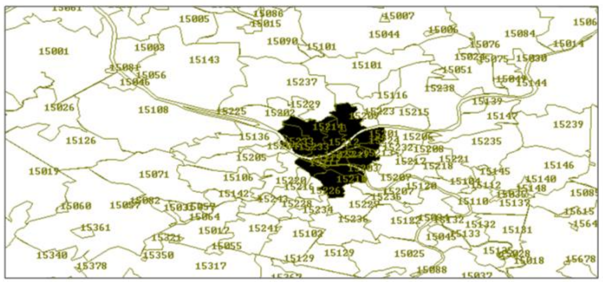
GFSS was better able to capture the shape of irregular clusters.

True outbreak region | MBSS (p = 1) | GFSS (p = 0.5)



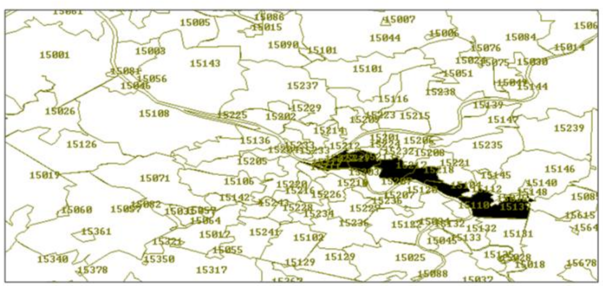
Generalized Fast Subset Sums

Optimization of the sparsity parameter p can substantially improve the detection performance of the GFSS approach.



Compact cluster:

Detection time minimized at p = 0.5;
spatial accuracy maximized at p = 0.7.

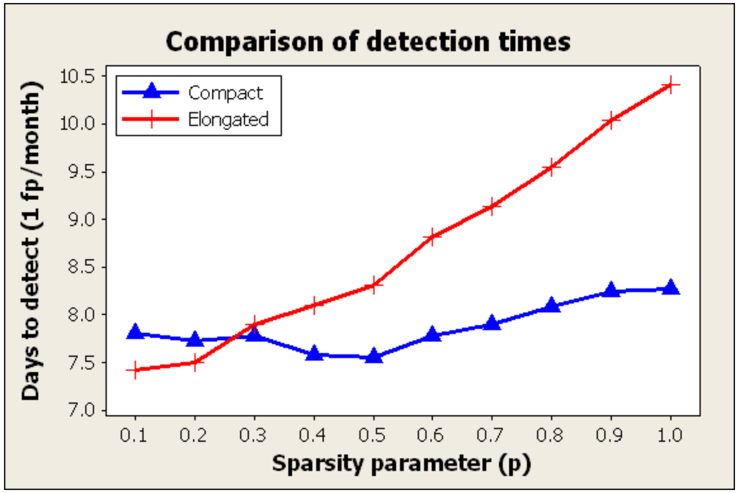


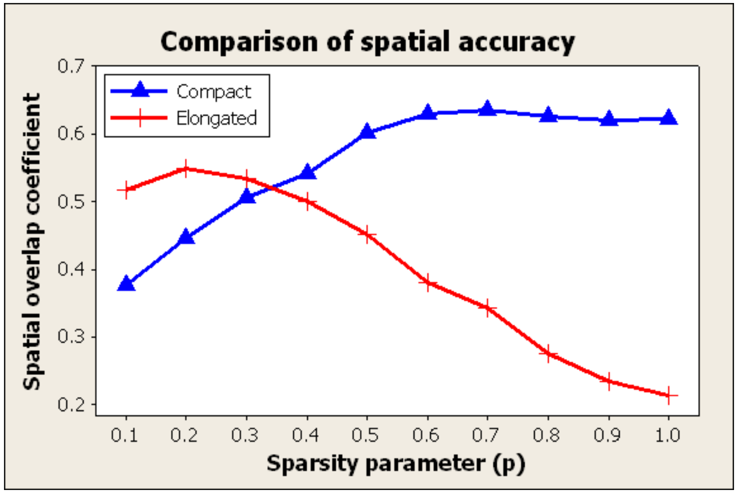
Highly elongated cluster:

Detection time minimized at p = 0.1;
spatial accuracy maximized at p = 0.2.

Generalized Fast Subset Sums

Optimization of the sparsity parameter p can substantially improve the detection performance of the GFSS approach.





For elongated clusters, p = 0.2 improves detection time by 0.8 days and spatial accuracy by ~10%, as compared to p = 0.5.

Conclusions

GFSS shares the essential advantages of MBSS: it can integrate information from **multiple data streams**, and can accurately distinguish between **multiple outbreak types**.

As compared to the MBSS method, GFSS substantially improves **accuracy** and **timeliness** of detection for elongated or irregular clusters, with similar performance for compact clusters.

While a naïve computation over the exponentially many subsets of the data is computationally infeasible, GFSS can **efficiently** and **exactly** compute the posterior probability map.

We can also **learn** the prior distributions over centers and neighborhoods and the sparsity parameter p for each event type using a small amount of training data. This enables us to better differentiate between multiple, similar types of outbreak.

References

- **D.B. Neill. Fast Bayesian scan statistics for multivariate event detection and visualization. *Statistics in Medicine,* 2010, in press.**
- D.B. Neill and G.F. Cooper. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning* 79: 261-282, 2010.
- M. Makatchev and D.B. Neill. Learning outbreak regions in Bayesian spatial scan statistics. *Proc. ICML Workshop on Machine Learning in Health Care Applications*, 2008.
- D.B. Neill. Incorporating learning into disease surveillance systems. *Advances in Disease Surveillance* 4: 107, 2007.
- D.B. Neill, A.W. Moore, and G.F. Cooper. A multivariate Bayesian scan statistic. *Advances in Disease Surveillance* 2: 60, 2007.
- D.B. Neill, A.W. Moore, and G.F. Cooper. A Bayesian spatial scan statistic. In *Advances in Neural Information Processing Systems* 18, 2006.