SAND2012-5809C

Spatial and Temporal Data Fusion for Biosurveillance

Karen Cheng, David Crary

Applied Research Associates, Inc.

Jaideep Ray, Cosmin Safta, Mahamudul Hasan

Sandia National Laboratories

Contact: Ms. Karen Cheng, kcheng@ara.com, 703-816-8886 x 138

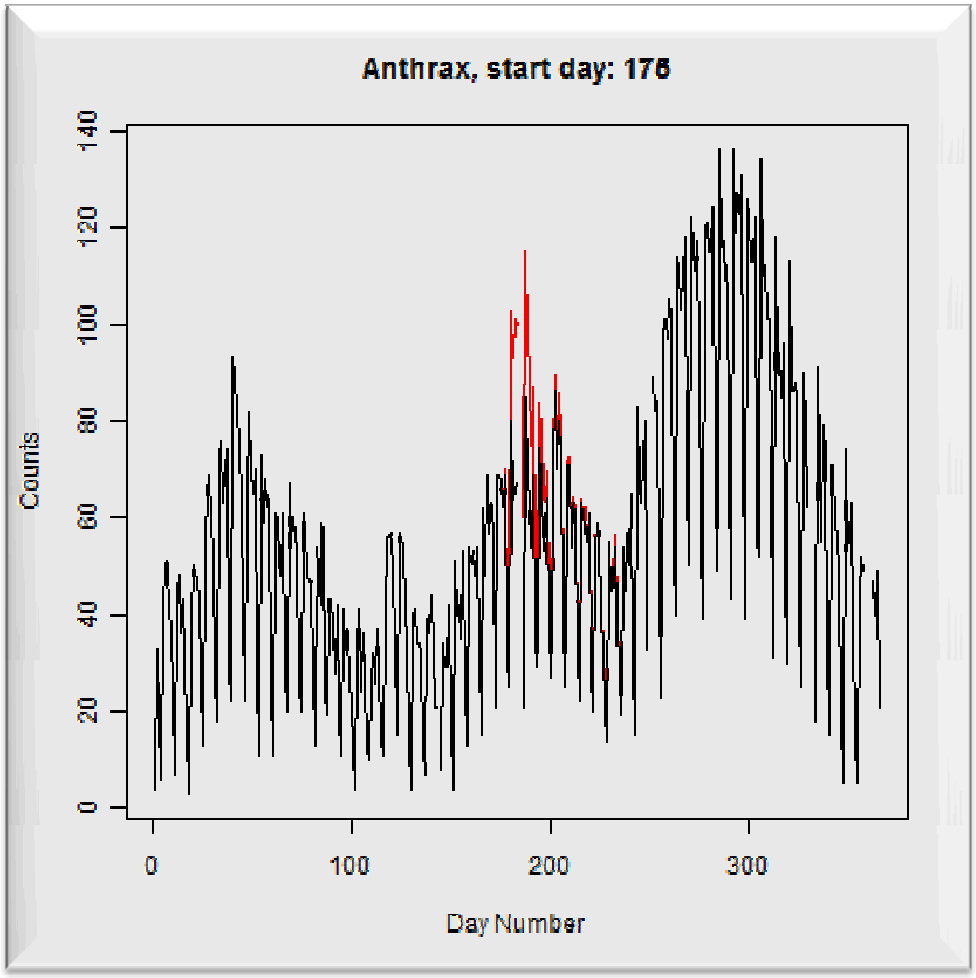
SAND 2012-5809C

Steps for Detecting, Characterizing, and Identifying an Outbreak from Syndromic Surveillance Data

- The components of the procedure are:
  - **Background Modeling/Outbreak Detection** from time-series data
    - Data contains the outbreak and background/endemic morbidity
  - **Extraction** of the outbreak from the background
    - Endemic component needs to be separated from the epidemic component
  - **Characterization** of the outbreak
    - estimation of index cases, time/rate of infection
  - **Identification** of the outbreak
    - What was the disease that caused it, given a few competing guesses

Previous Analysis with Purely Temporal Information

Simulated Anthrax Attack on Day 175



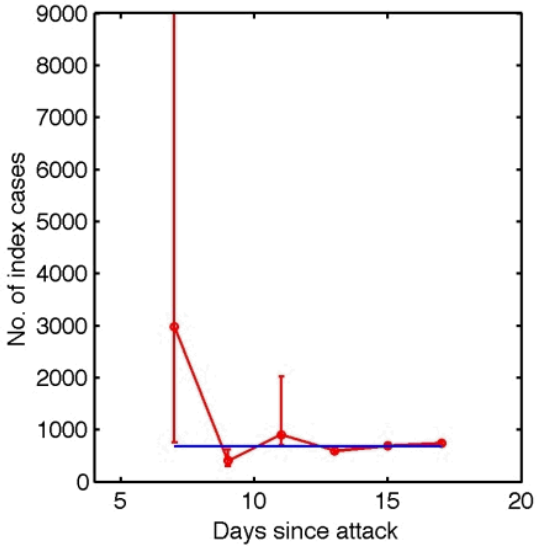
- Background: ILI ICD-9 codes from Miami data
- Red Line: Calculated anthrax outbreak from Wilkening A2 model, plus visit delay; 500 index cases

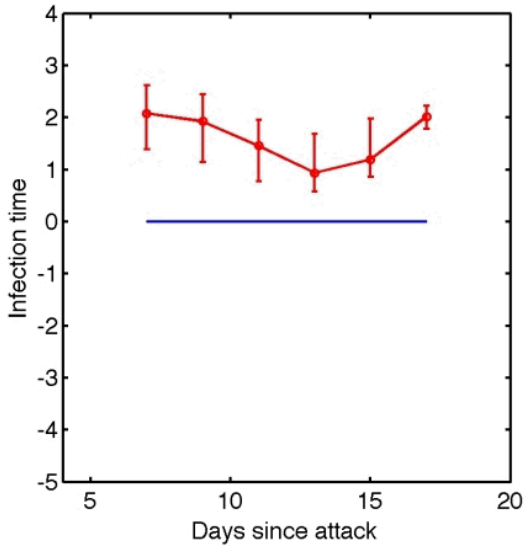
We get an alarm on day 180.

We Used Bayesian Techniques to Characterize the Outbreak

- We formulate the estimation as a Bayesian inverse problem
  - Predicated on the extracted epidemic data
- Allows one to use bounds / prior beliefs regarding the value of the parameters
  - We assumed that index cases ranged between 100-10,000
- Solved using an adaptive Markov Chain Monte Carlo sampler
  - All parameters estimated as probability density functions (PDF)
  - Used autocorrelation analysis to determine “convergence” of the Markov chain

How Small An Outbreak Can We Characterize?





Number of index cases and time of attack for an anthrax outbreak with 680 index cases. True values indicated in blue

- Tested on simulated anthrax epidemic of various sizes
- Could estimate N_{index} and τ for the attack >= 680 infected cases

Identification of Outbreak Using Syndromic Surveillance Data

- ***What if the identity of the disease was unknown?***
  - How would you reconstruct the outbreak?
- ***Approach:***
  - Shortlist candidate diseases, based on symptoms
  - Characterize outbreak with each disease model
  - Using the distribution of epidemic model parameters, forecast the epidemic
    - Compare with observed data (posterior predictive tests) and calculate goodness-of-fit
- Identified ensemble metrics that calculate goodness-of-fit:
  - Best fit model identified by "voting" the metrics
  - Metrics: CRPS, MAE, IS90, IS80, IS50
- Correctly identified the causative agent:
  - With 4 competing diseases with similar incubations and flu-like symptoms (anthrax, plague, smallpox and flu)

Days of data	Plague	Anthrax	Flu	Smallpox
5	0/5	0/5	0/5	5/5 (CRPS, MAE, Isxy)
7	0/5	0/5	0/5	5/5
9	0/5	0/5	0/5	5/5
11	0/5	0/5	0/5	5/5
13	0/5	0/5	0/5	5/5

Scores of each disease model, when fitted to early-epoch Camp Custer data. Smallpox is correctly selected with 5 days of data. Metrics supporting a disease model are mentioned in parenthesis.

Competing anthrax, flu, smallpox and plague models on smallpox data
- Smallpox correctly identified in 5 days
- Corroborated over next 8 days

Initial Spatio-temporal Analysis - Introduction

- Syndromic surveillance data is spatio-temporal
  - We generally have the ZIP-codes of infected people
- To date, we've aggregated syndromic surveillance data up to city levels and performed purely temporal characterization
- Can spatial data help us do better?
- Contemporary Methods
  - Take the available data and cluster it; will provide a good region to concentrate resource allocation
  - As more data becomes available, and clusters widen / increase in number, widen your area of interest (evidence-based approach)
  - Limitation: lacks understanding of the source incident, timeliness for planning
- Conjecture : Can we infer the future region of infection (where others will turn up sick) with sparse data?

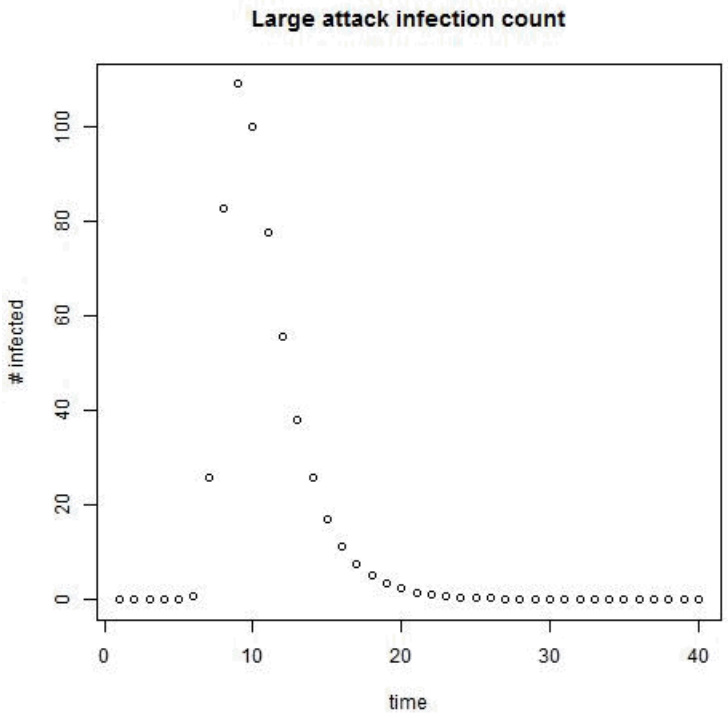
Approach

- The key to estimating clusters of infected people is to characterize the attack
  - Location, size and time, inferred probabilistically
  - And then use a dispersion model + epidemic model to identify where the incubating and susceptible people are (we already know the symptomatic ones)
- How? The model
  - Use a dispersion model to "spread" an aerosol & infect people with different doses
    - Inputs: location of release, amount of release
    - Gaussian puff model – D. B. Turner, *Workbook of Atmospheric Dispersion Estimates: An Introduction to Dispersion Modeling.*
  - Use an epidemic model (say, for anthrax) to predict the evolution of the disease, given infected people with varying doses
    - Pick anthrax, and use Wilkening's A1 model for incubation period
    - Inputs: time of infection, # of infected people and their dosages.

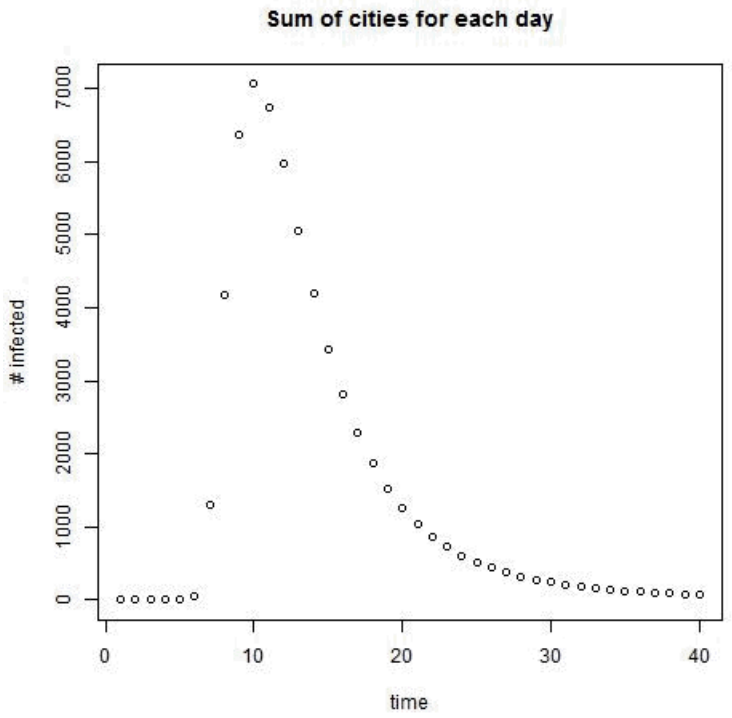
Approach (cont.)

- Inverse problem
  - Data: # of symptomatic people, per day, per zip-code (whose location is known)
  - To infer: (x, y, z) location of release point, Q, the # of spores released, t the number of days before 1st symptoms, when the people were infected
  - Assume: Gaussian noise in observations
- Solution:
  - Use MCMC to create posterior distributions for (x, y, z, $\log_{10}(Q)$, t)
- Tests
  - Test with synthetic data, generated using Wilkening A1 model
    - With sufficient data, we should infer the true release point
  - Can small attacks be inferred? How well?
  - Test with synthetic data, generated using Wilkening's A2 model
    - Even with infinite data we will not infer back the true parameters
    - But will we come close? How close?

Case I – Big Attack with No Model Mismatch



Epidemic curve for a chosen zip-code

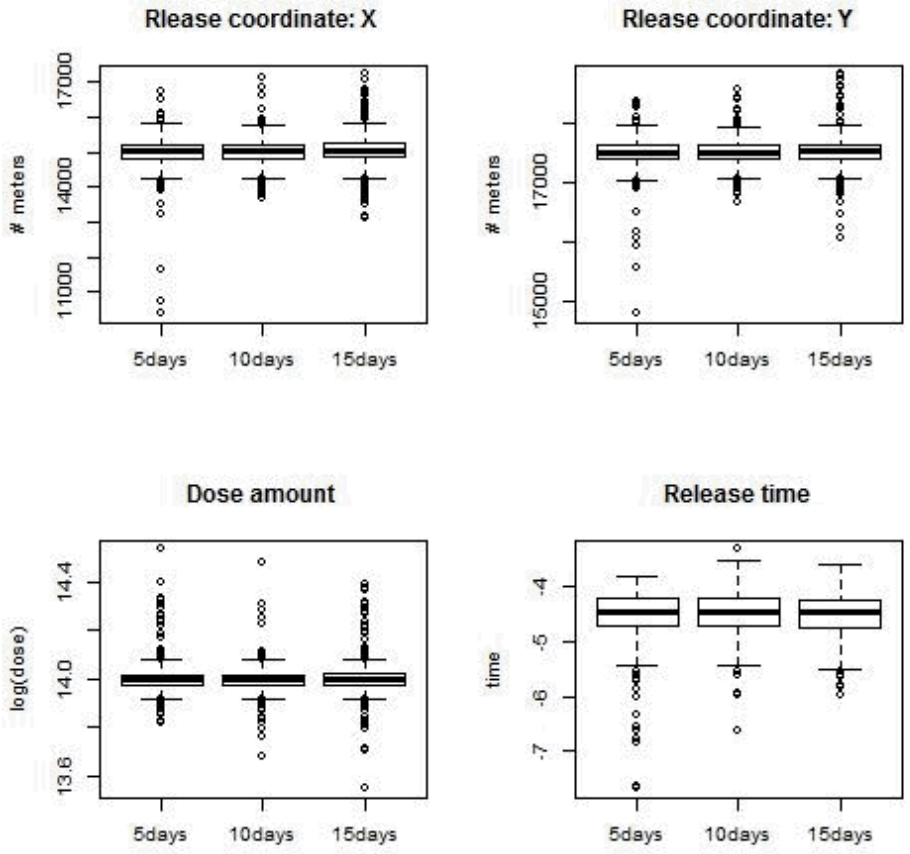


Epidemic curve for the entire city

- 50 km X 50 km city, divided into 1 km x 1km grid-cells
- Left – epidemic curve in a grid-cell
- Right – epidemic curve summed over all grid-cells

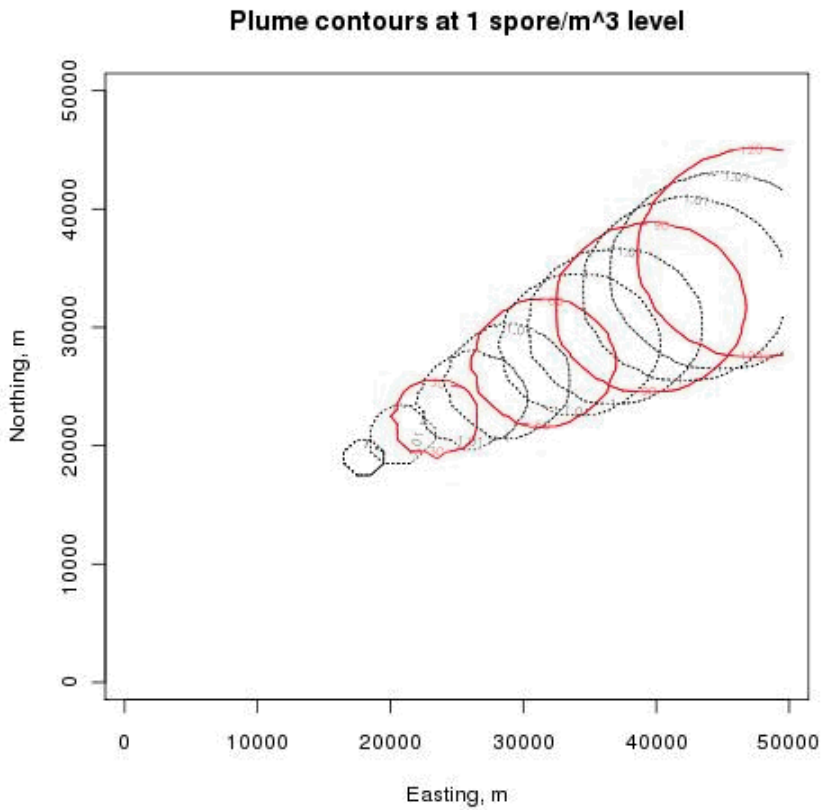
Inferred Location, Quantity and Time of Release

- For large attack, even 5 days of data is good enough
- True values:
  - X : 15,000 m
  - Y : 17,500 m
  - Log_{10}(Dose) = 14
  - Time = -5 days

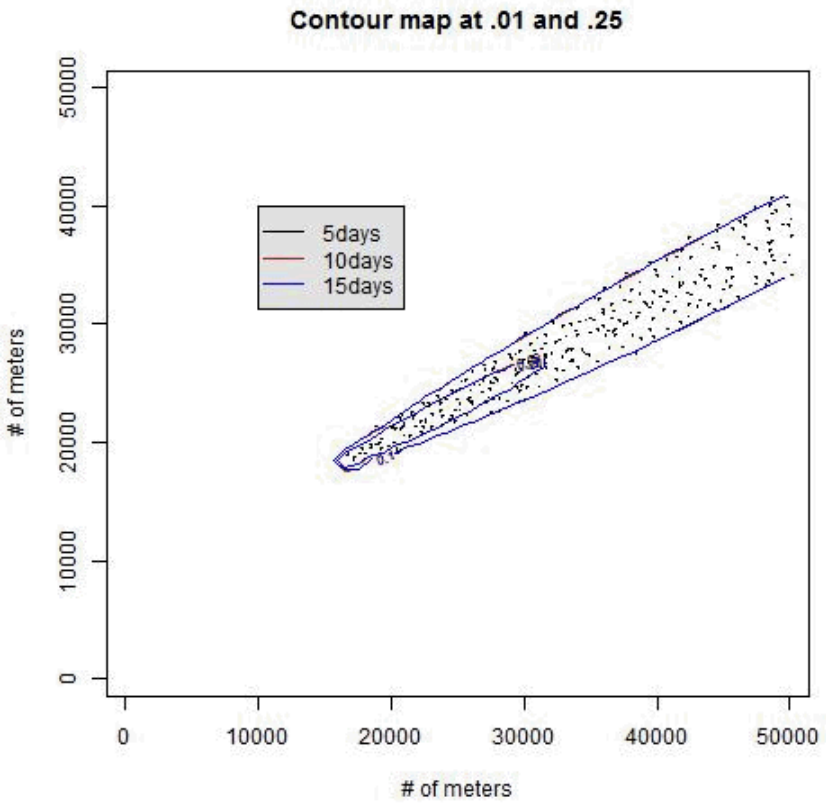


Inferred values of release location (X, Y), release size (log_{10}(Q)) and release time. True values [15,000; 17,500; 14, -5]

Clusters – Observed and Predicted



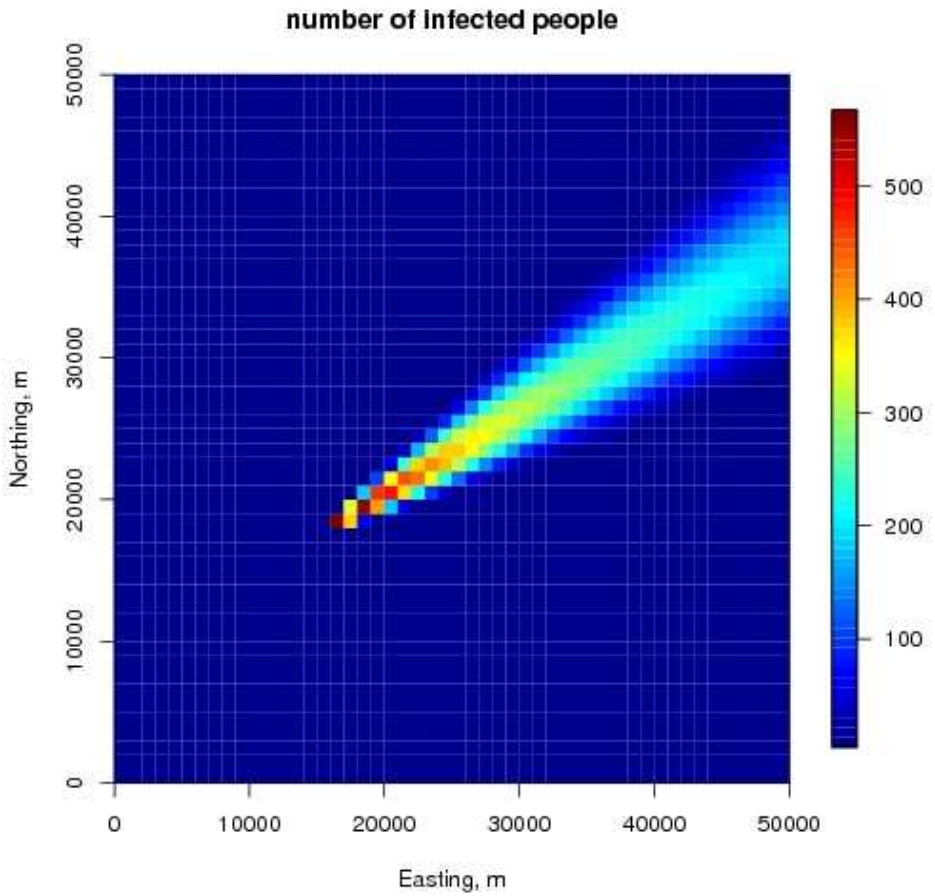
Inferred contours of spore concentration. Red contours are at 30 min intervals.



Contours show regions where 1% (outer) and 25% (inner) of the population are infected as a result of the release. Dots are individuals reporting.

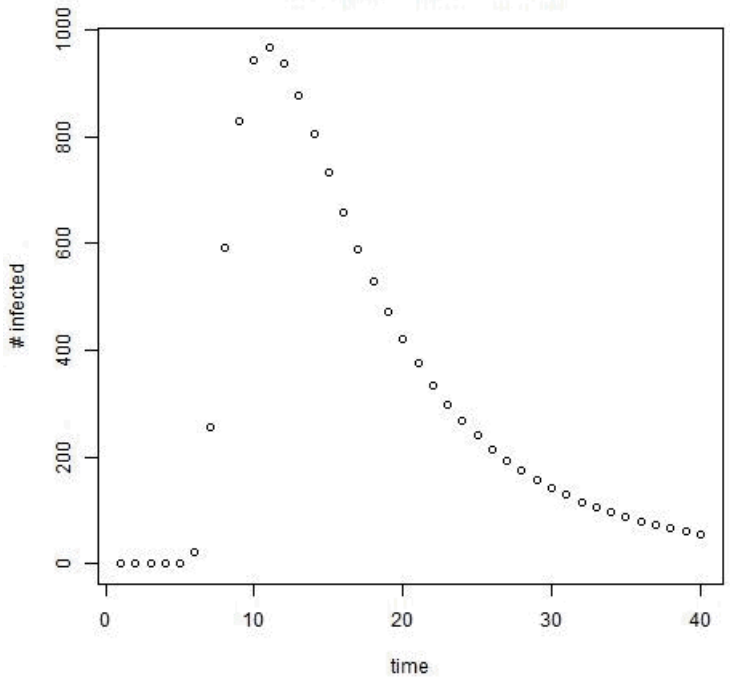
Estimated Distribution of Infected People

- Another, colorful view
- Infected people concentrated along centerline of plume

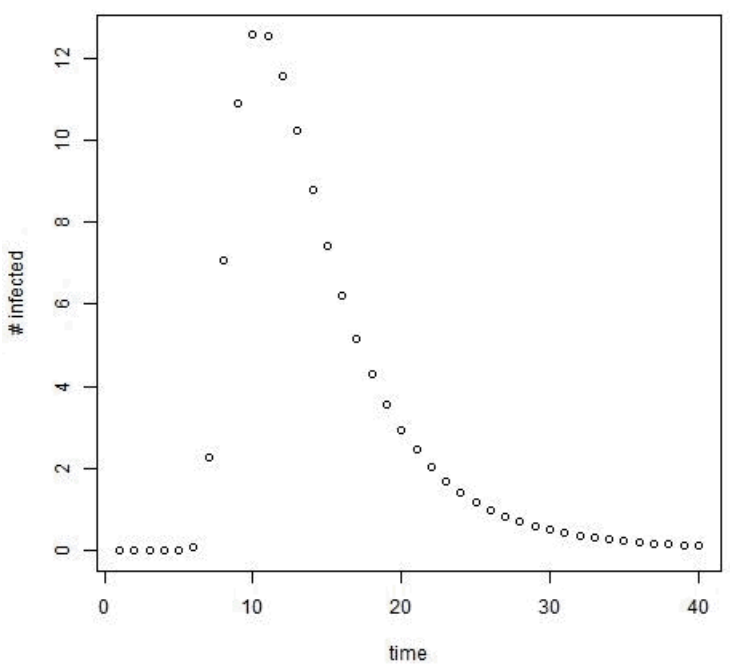


Case II – Small Attack with no Model Mismatch

Epidemic curve for the entire city



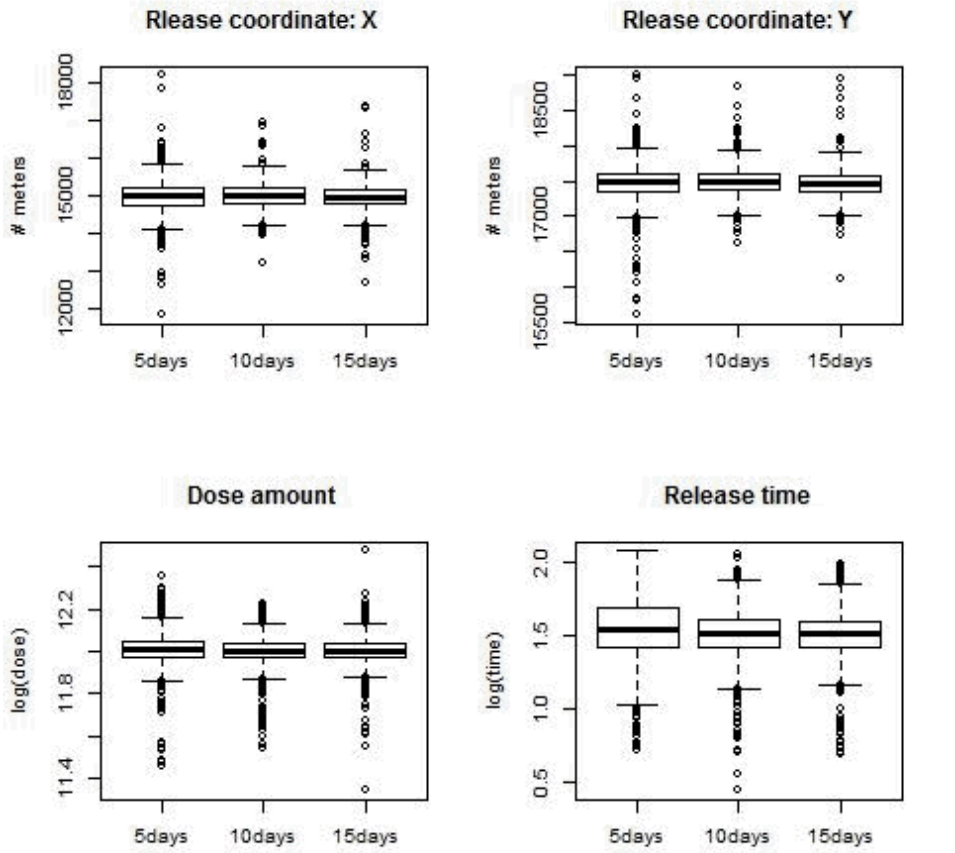
Epidemic curve for a chosen zip-code



- 50 km X 50 km city, divided into 1 km x 1km grid-cells
- Left – epidemic curve in a grid-cell
- Right – epidemic curve summed over all grid-cells

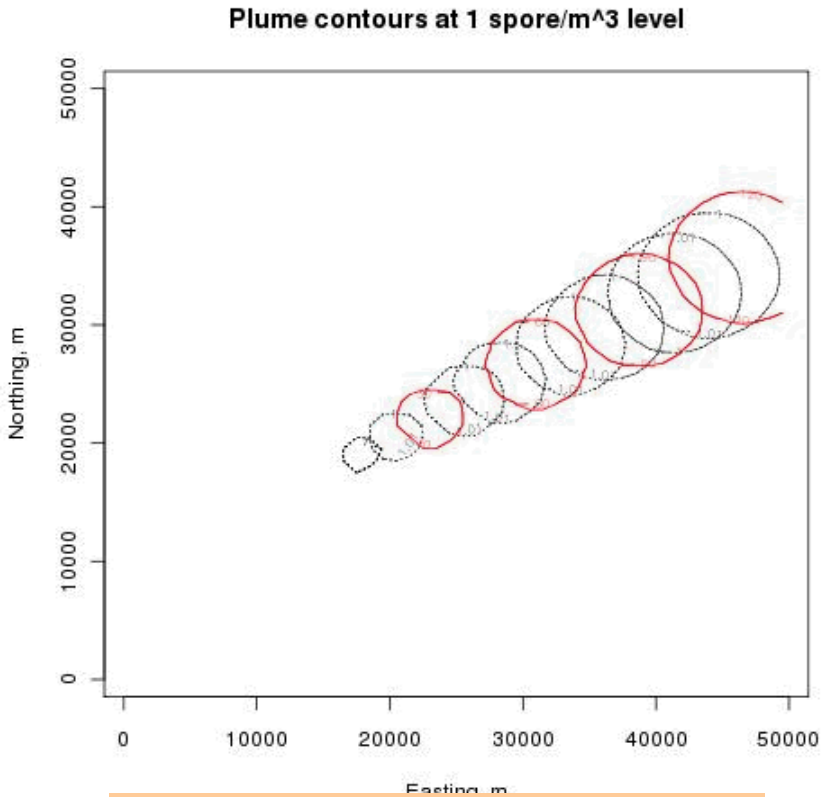
Inference of Release Parameters

- Again easy to infer
- True values:
  - X : 15,000 m
  - Y : 17,500 m
  - Log_{10}(Dose) = 14
  - Time = -5 days

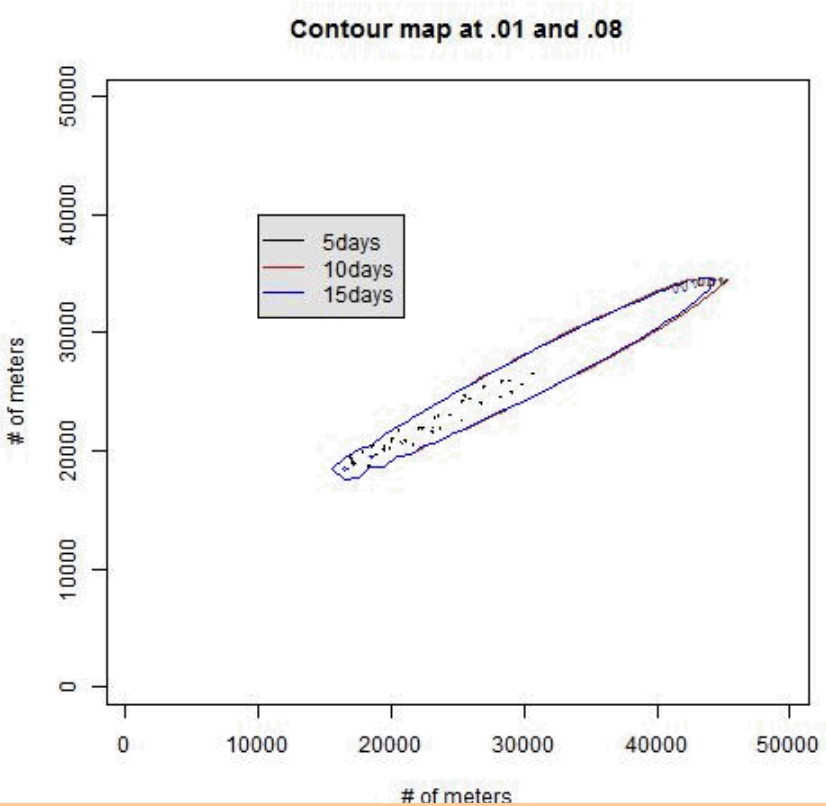


Inferred values of release location (X, Y), release size (log_{10}(Q)) and release time. True values [15,000; 17,500; 12, -5]

Contours – Observed and Predicted



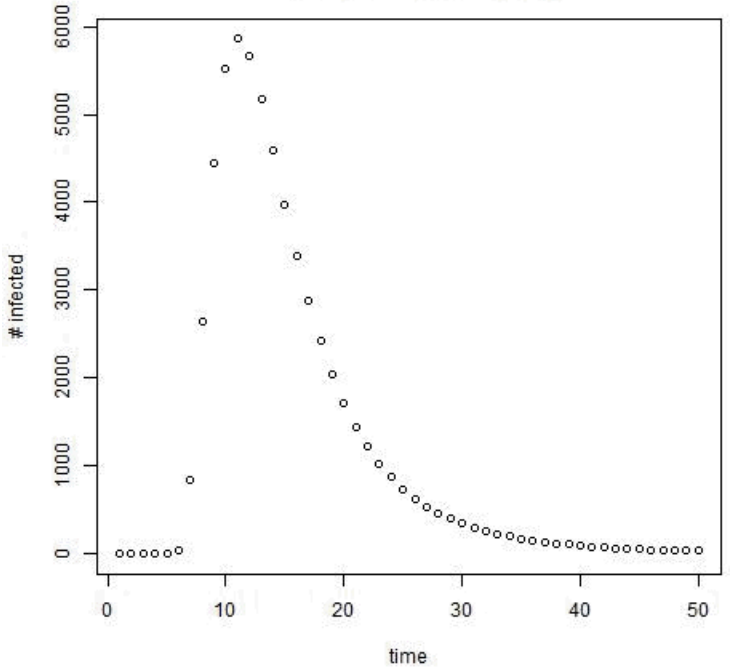
Inferred contours of spore concentration. Red contours are at 30 min intervals.



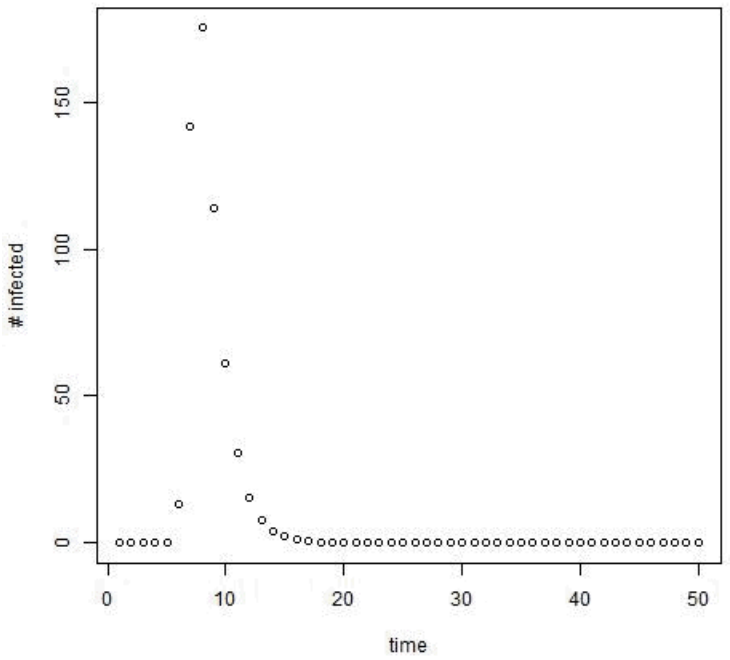
Contours show regions where 1% (outer) and 25% (inner) of the population are infected as a result of the release. Dots are individuals reporting.

Case III – Inference under Model Mismatch

Epidemic curve for the entire city



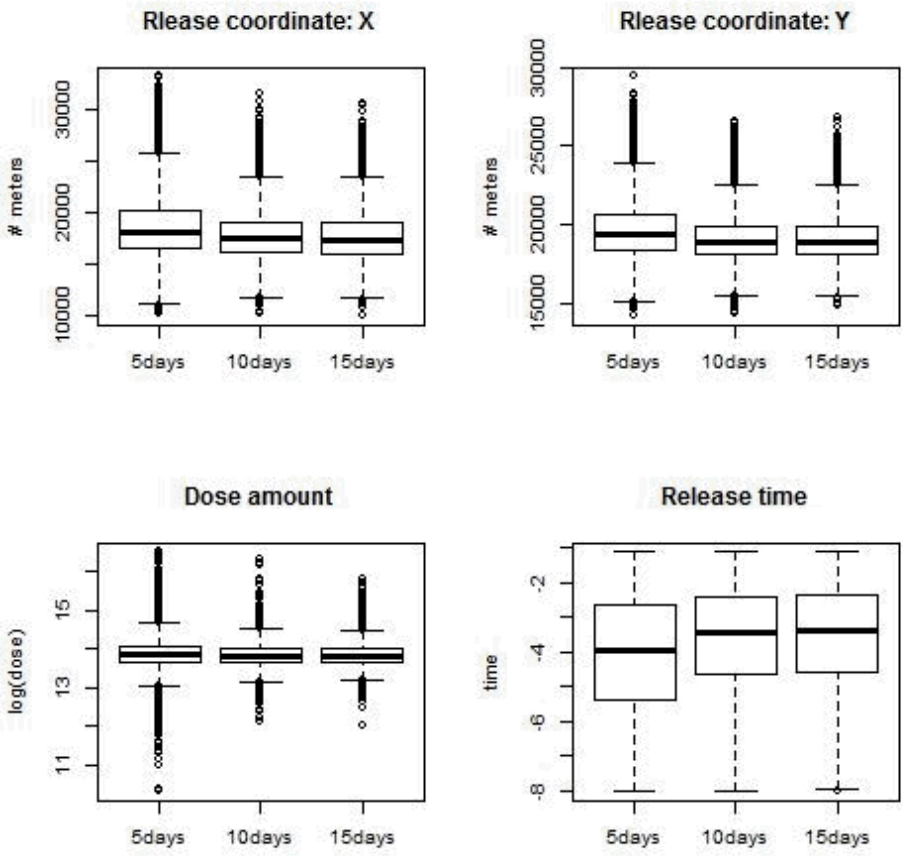
Epidemic curve for a chosen zip-code



- 50 km X 50 km city, divided into 1 km x 1km grid-cells
- Left – epidemic curve in a grid-cell
- Right – epidemic curve summed over all grid-cells

Inference of Release Parameters

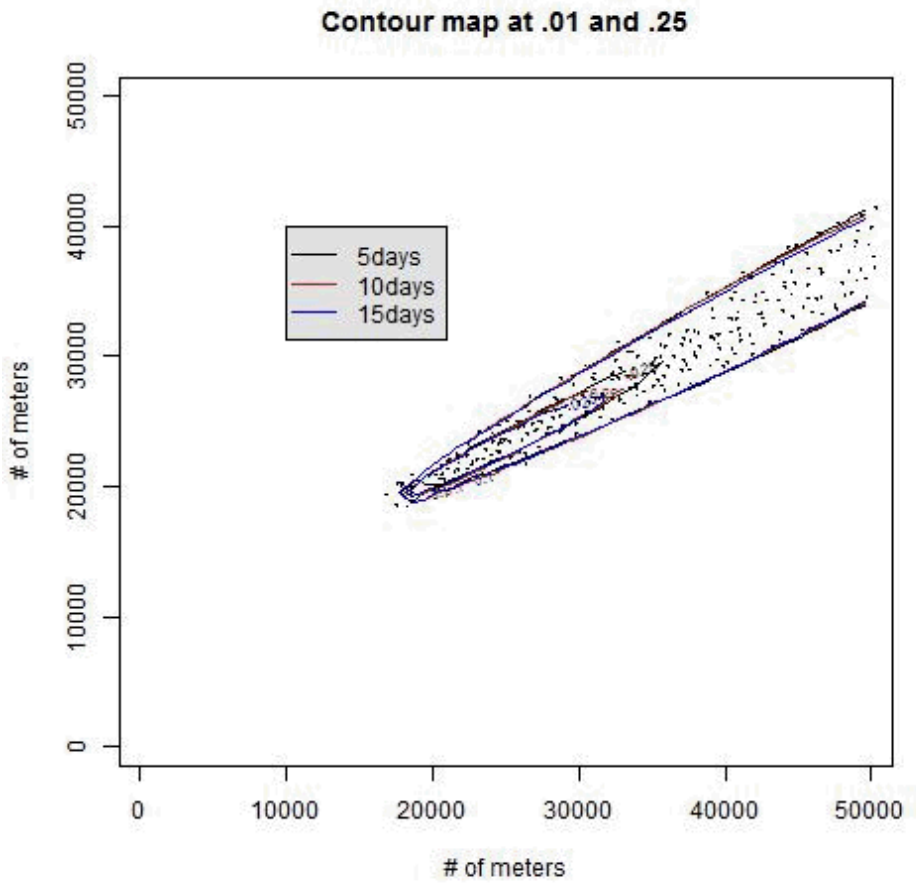
- Locations inferred wrongly – but by about 2 grid-cells (2 km)
- Underestimated release quantity
- Bigger uncertainties in time
- No improvement with addition of data (beyond 5 days)



Inferred values of release location (X, Y), release size ($\log_{10}(Q)$) and release time. True values [15,000; 17,500; 14, -5]

Contours – Observed and Predicted

Clustering still OK even with model mismatch

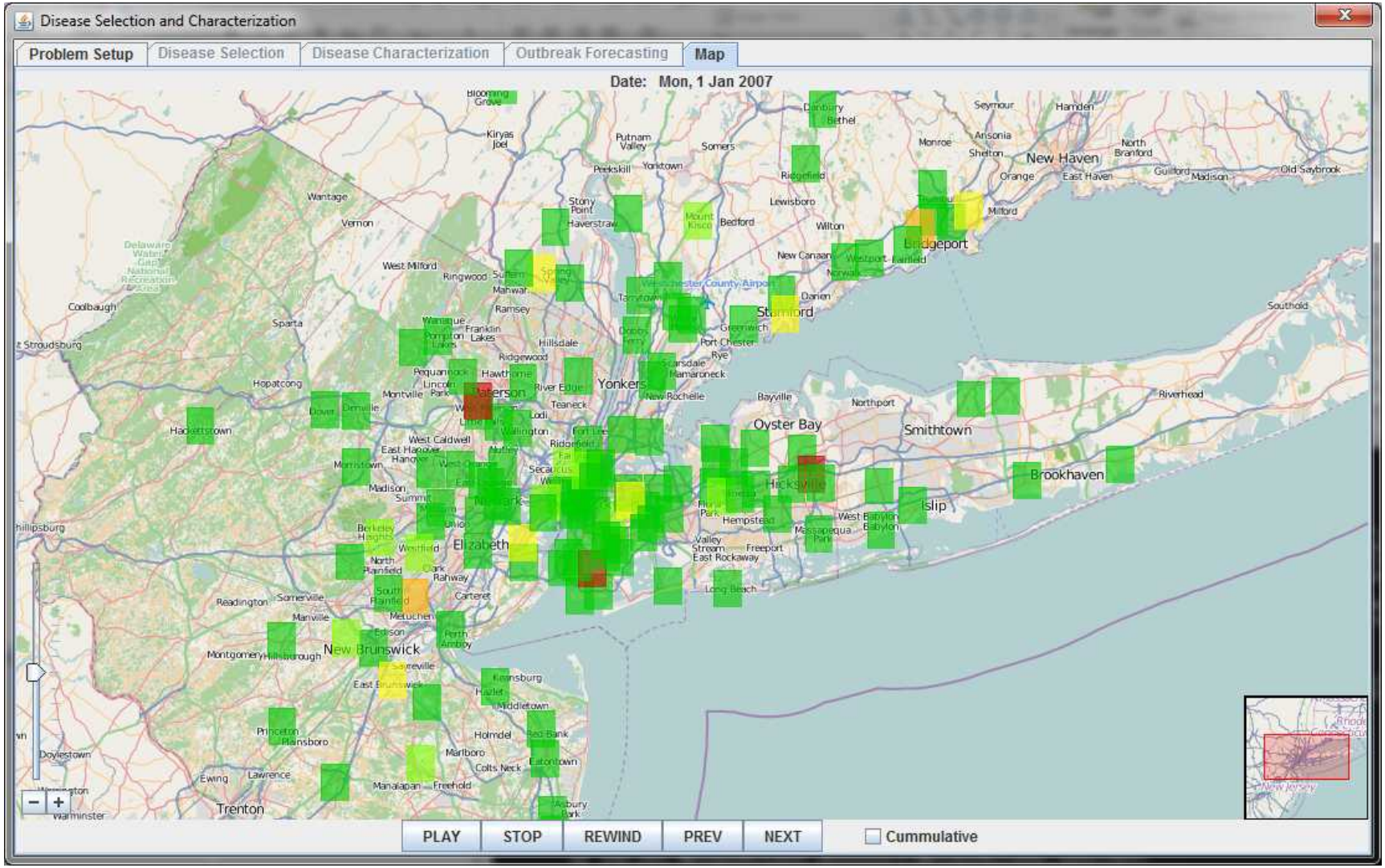


Contours show regions where 1% (outer) and 25% (inner) of the population are infected as a result of the release. Dots are individuals reporting.

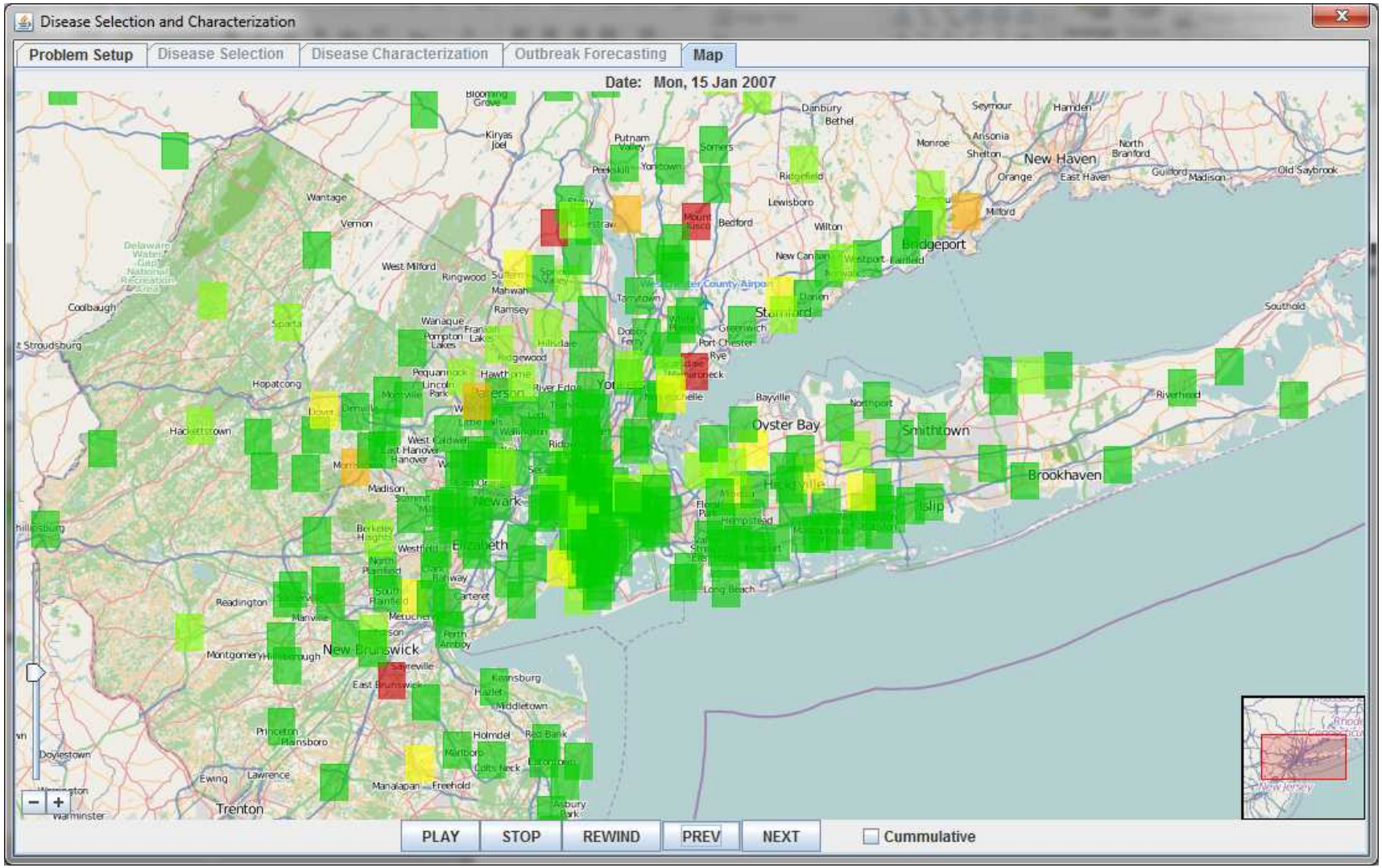
Temporal-spatio Visualization Prototype

- Pure visualization alone is very useful for understanding outbreaks
- Prototype "Heat Map" of reports by zip code
  - Color based on number of events
  - Current day or cumulative counts
  - Animates changes in "playback" mode through time
- Future Enhancements in progress
  - Add source term estimation
  - Add prediction capabilities for contagious diseases
    - Will use the underlying model within our Kalman Filter Anomaly Detector to future predict case counts
    - Use random field models to disperse cases

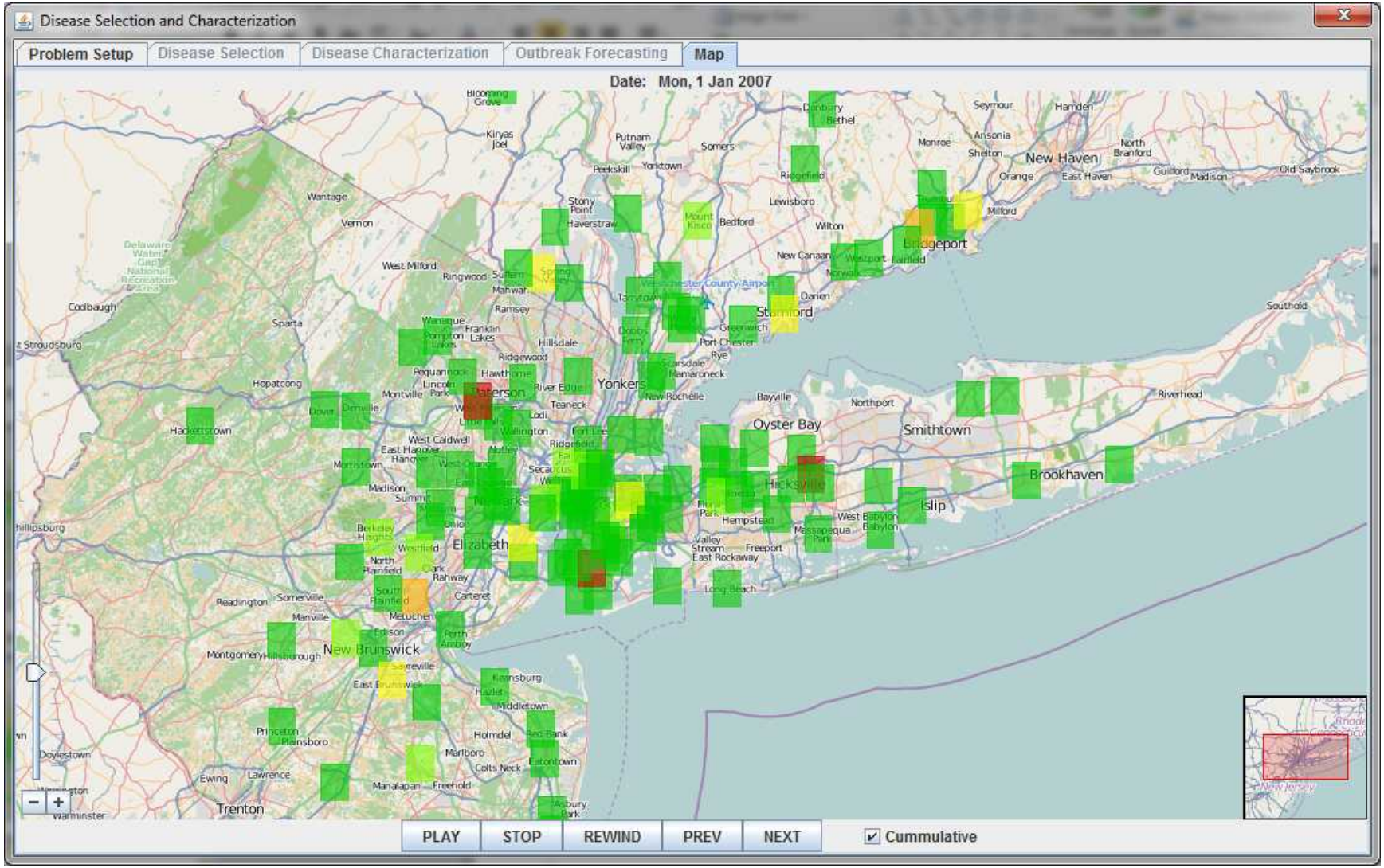
Daily Report Heat Map



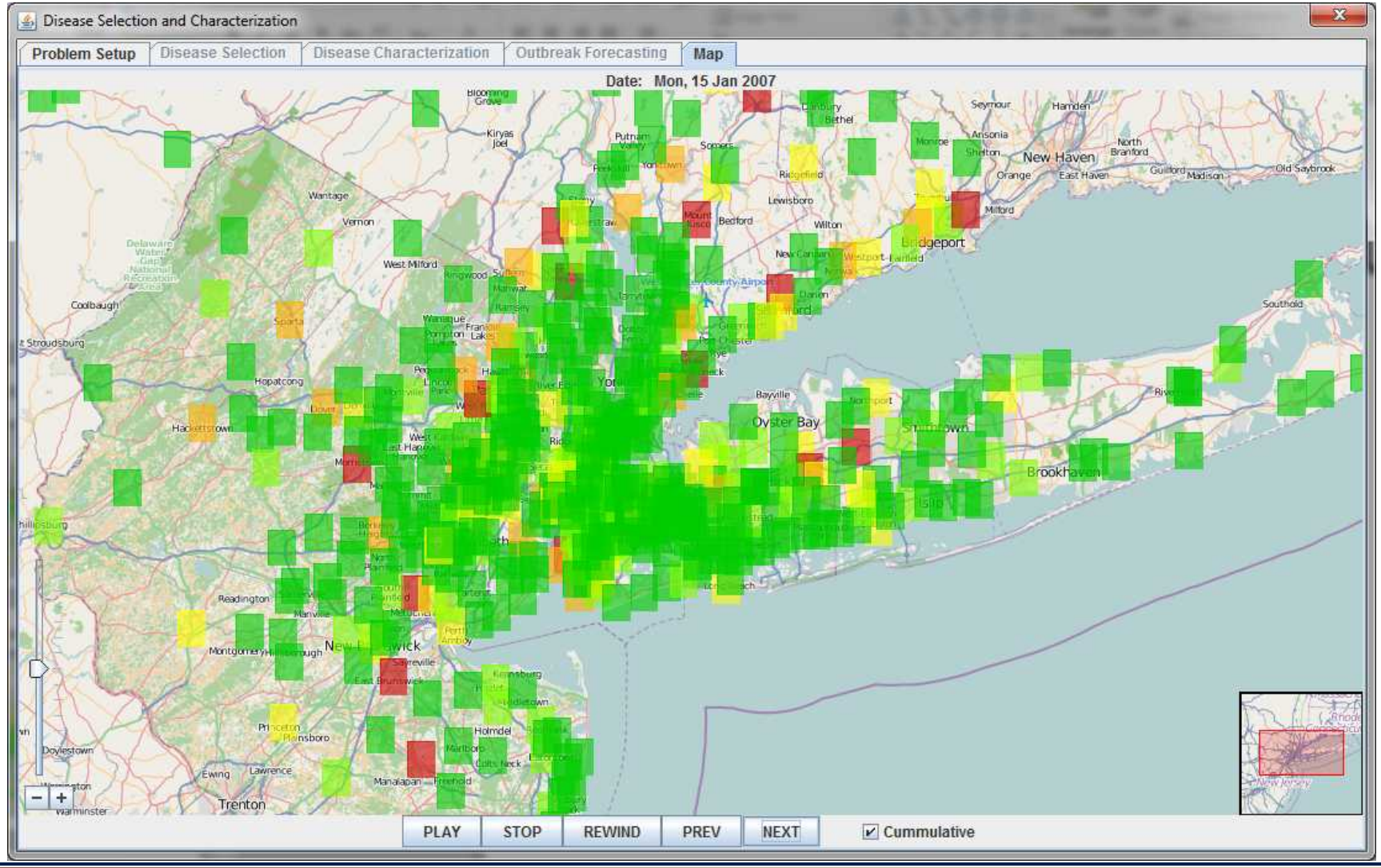
Daily Report Heat Map



Cumulative Report Heat Map



Cumulative Report Heat Map



Acknowledgements

This work is funded by the Defense Threat Reduction Agency (DTRA)
under contract HDTRA1-09-C-0034
Ms. Nancy Nurthen at DTRA is the Program Manager.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000."