Spatial and Temporal Data Fusion for Biosurveillance

Karen Cheng, David Crary

Applied Research Associates, Inc.

Jaideep Ray, Cosmin Safta, Mahmudul Hasan

Sandia National Laboratories

Contact: Ms. Karen Cheng, kcheng@ara.com, *571-814-2411*

Early Event Detection and Characterization

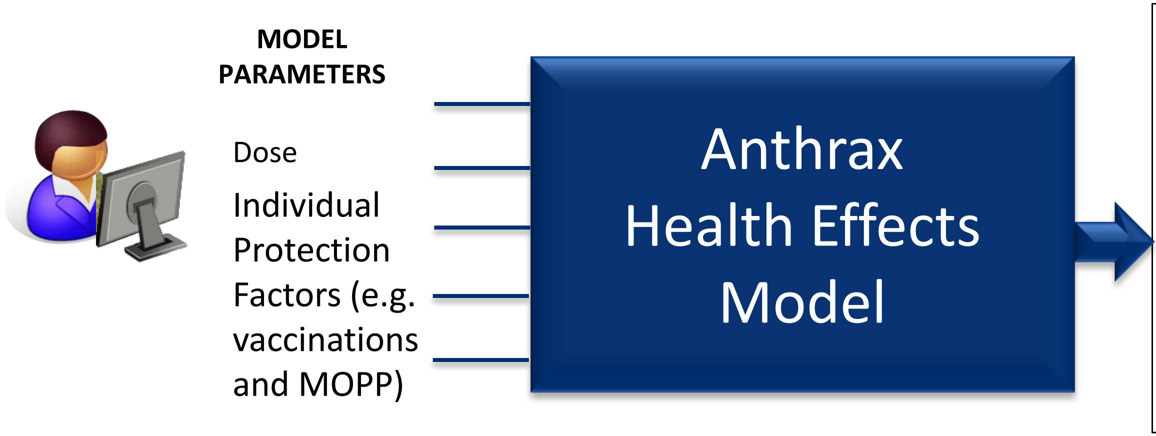
- Early on in an outbreak (malicious or naturally-occurring) we will probably not know what the characteristics of the outbreak are
- What we do have today (e.g. hospital admission and discharge data) is:
  - Temporal data (e.g. number of hospital admissions on a daily basis)
  - Spatial data (e.g. the zip codes of the patients)
- We have focused on analyzing this data (available in hospitals or biosurveillance systems) to
  - **Characterize** the event
  - **Predict** the event
- My previous talk focused on **temporal** characterization. This talk emphasizes **spatial** characterization.

Characterization

- Both **temporal** and **spatial** characterization rely on

INFERENCE

- What is inference?
- In deliberate planning (what-if scenario analysis that assesses the damage of a theoretical event), analysts use **health effects/disease models.**
- The analyst sets the **parameters** of these **models** as he desires to assess worst case scenarios and perform medical planning

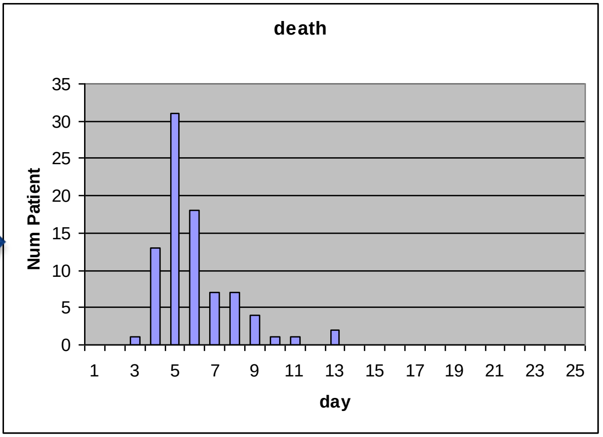


MODEL PARAMETERS

Dose

Individual Protection Factors (e.g. vaccinations and MOPP)

Anthrax Health Effects Model



What is Inference?

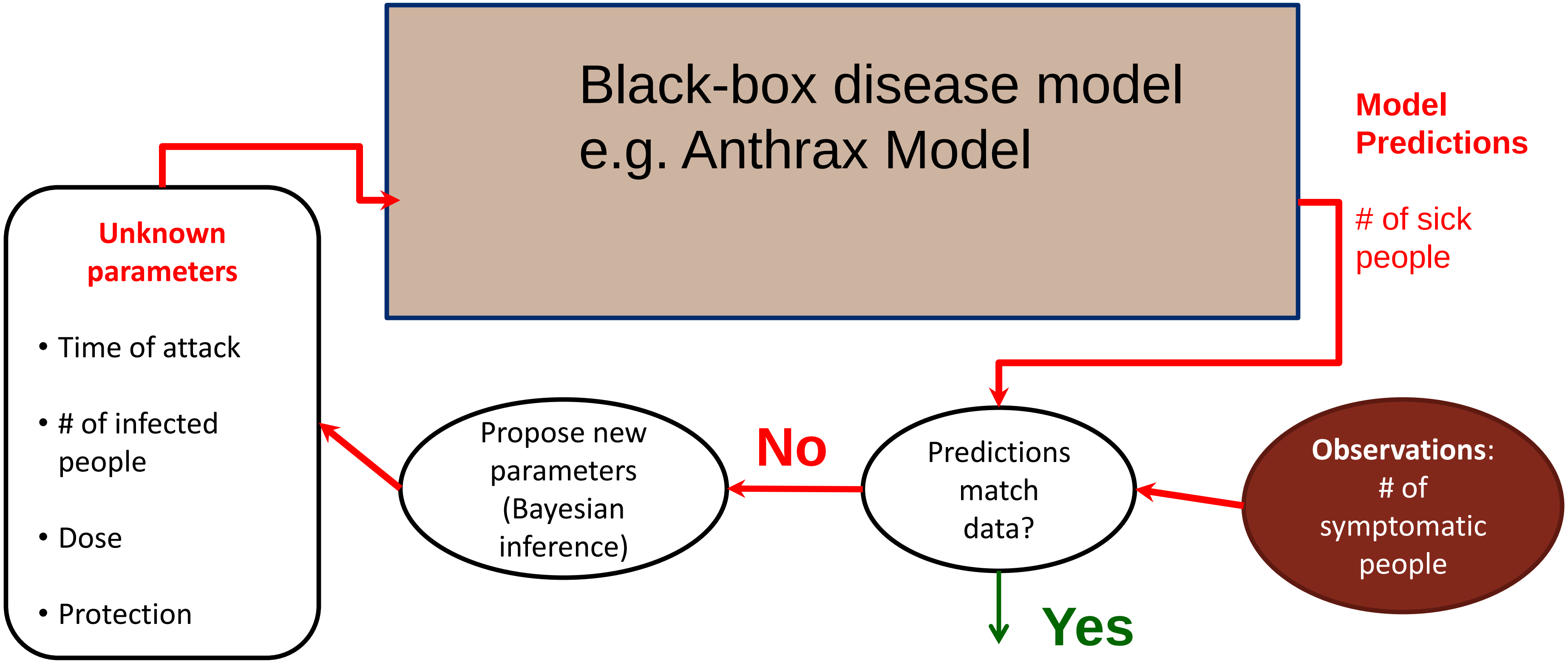
- In real-life situations (crisis response situations), early on, we have little understanding of what the event is.
  - All we have is data (usually can get spatial and temporal data) that represents some initial stage of the epidemic
- How can we do prediction?
- Answer: use the same models analysts use in **deliberate planning** for **crisis response planning**
- **Inference** is a technique that allows us to fit a particular model's (e.g. Plume Dispersion model's) parameters to the live data

Inference allows us to apply existing models to predict real-time crisis situations.
Prediction allows us to implement medical countermeasures and SAVE LIVES.

We Use Bayesian Techniques to Perform Inference to Characterize the Outbreak

- From Dr. Nicole Rosenzweig's talk yesterday
  - "decision makers make unambiguous decisions on very ambiguous data". What do we do about this?
- Bayesian techniques allow us to provide **confidence intervals** around our inferences and predictions (e.g. on a daily basis)
- Bayesian techniques infer the parameters of an outbreak model from the outbreak data available.
  - We formulate the estimation as a statistical inverse problem
    - You are given the "answer", so what caused it?
- Solved using an adaptive Markov Chain Monte Carlo sampler
  - All parameters estimated as probability density functions (PDF)

Inference – Fitting Models to Data: Disease Model



Black-box disease model
e.g. Anthrax Model

Model Predictions

# of sick people

Unknown parameters

- Time of attack
- # of infected people
- Dose
- Protection

Propose new parameters (Bayesian inference)

No

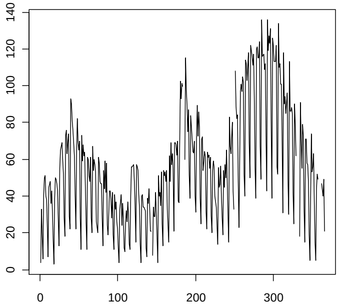
Predictions match data?

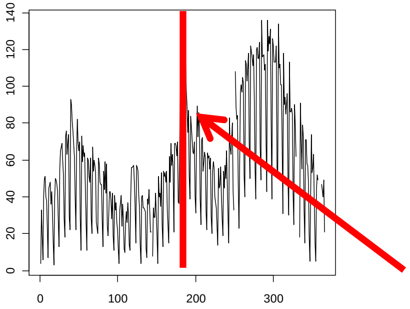
Observations:
# of symptomatic people

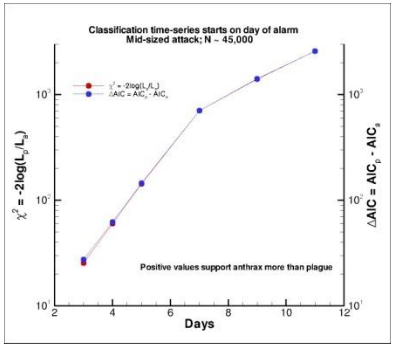
Yes

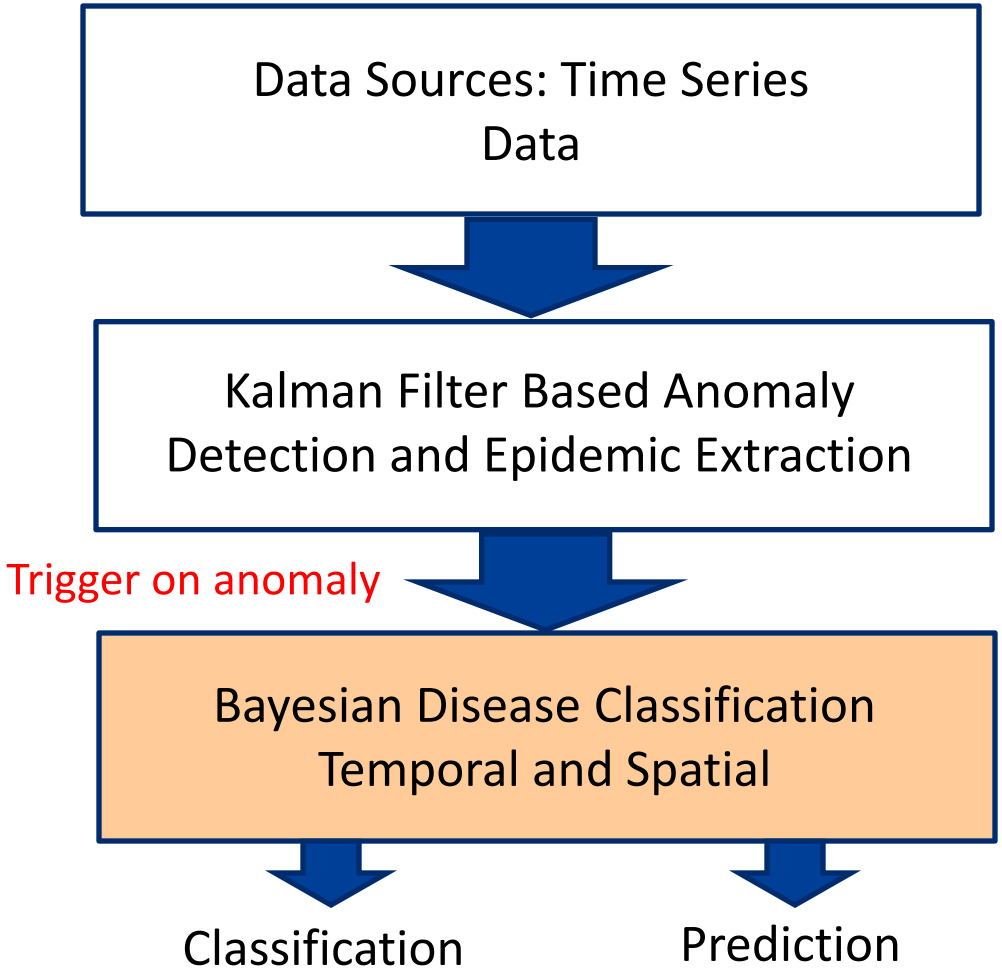
Save; Probable attack scenario

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









Data Sources: Time Series Data

Kalman Filter Based Anomaly Detection and Epidemic Extraction

Trigger on anomaly

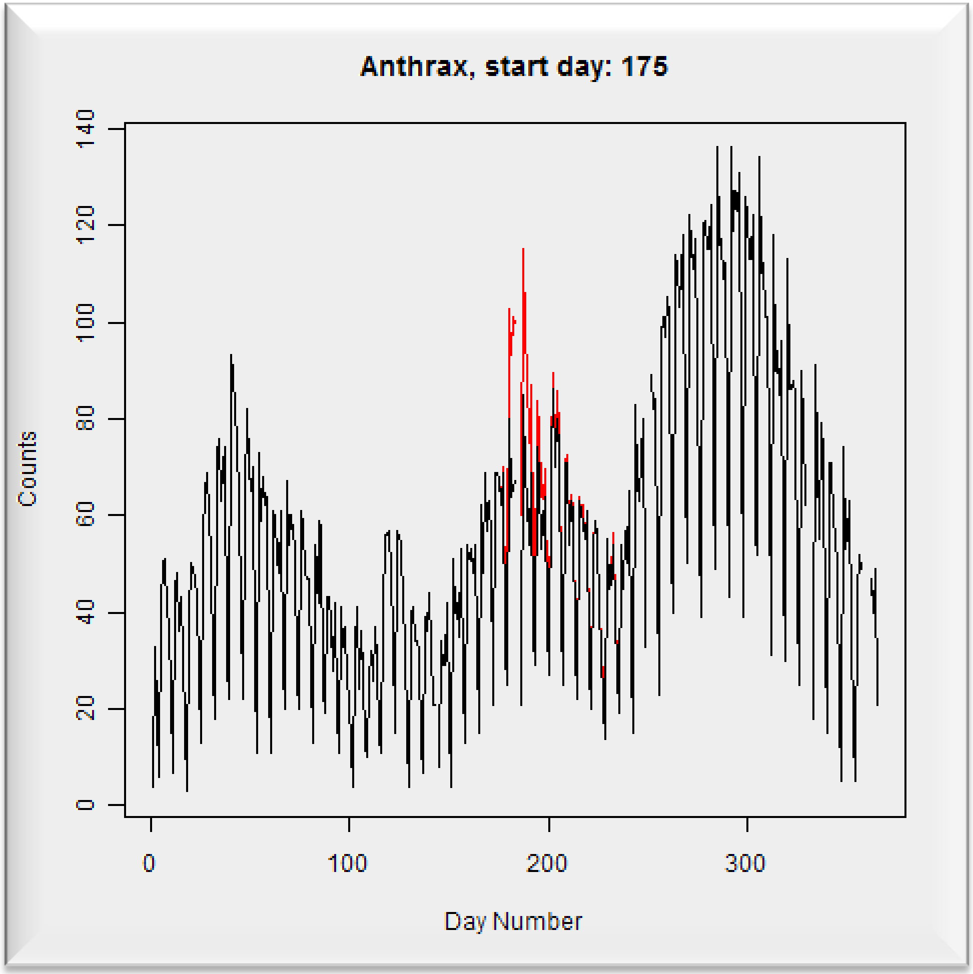
Bayesian Disease Classification Temporal and Spatial

Classification

Prediction

Previous Analysis with Purely Temporal Information

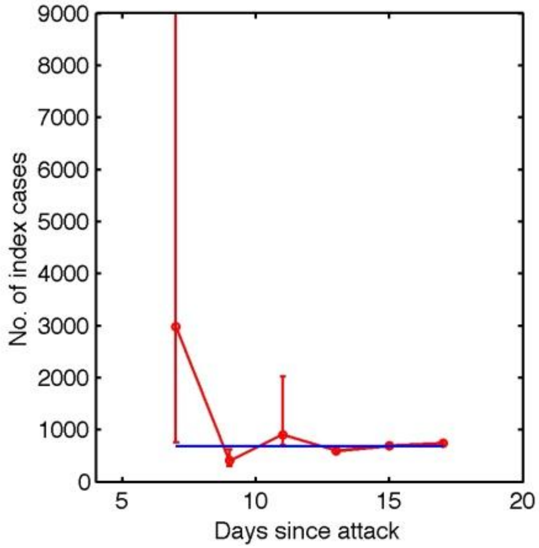
Simulated Anthrax Attack on Day 175



- 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.

How Small An Outbreak Can We Characterize?



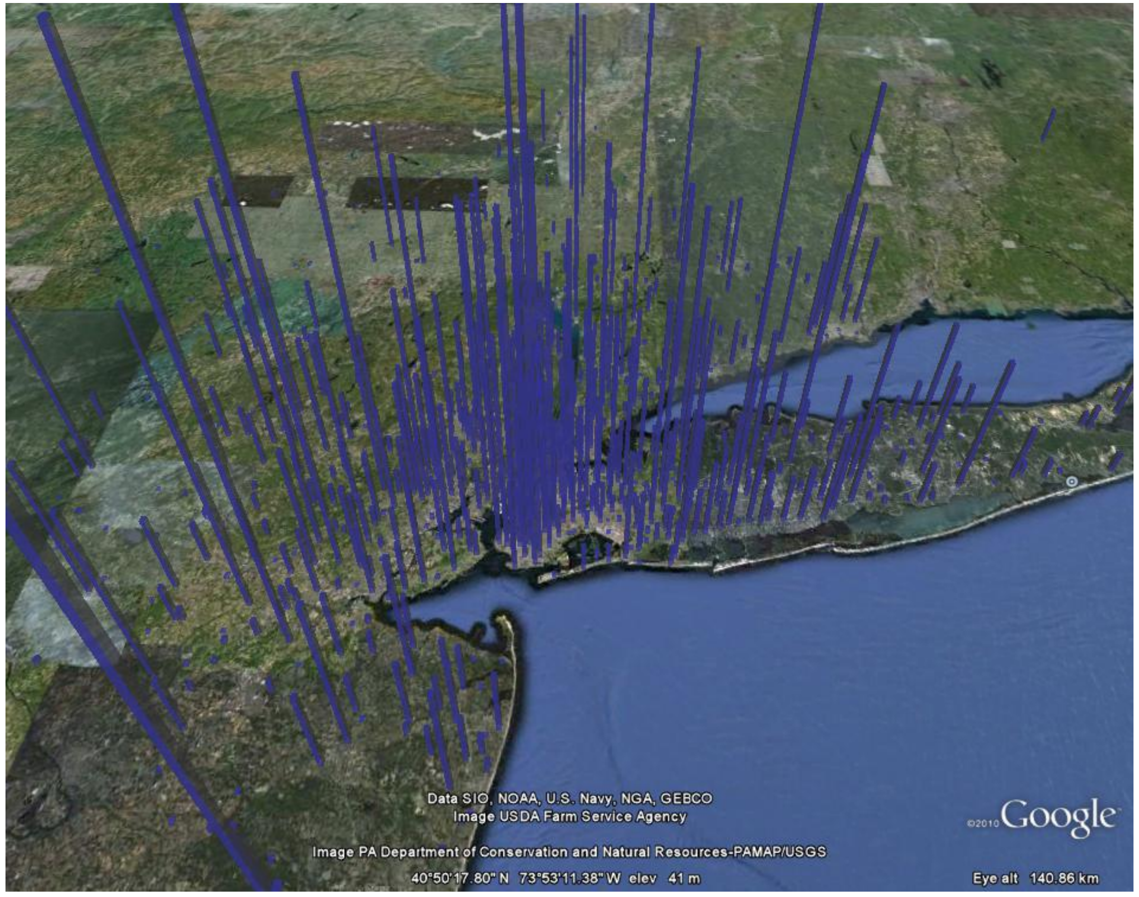
Number of index cases and time of attack for an anthrax outbreak with 680 index cases. True values indicated in blue

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

Initial Spatio-temporal Analysis - Introduction

- Syndromic surveillance data is spatio-temporal
  - We generally have the ZIP-codes of infected people
- Concept: Spatial data is a rich and very important source of information for disease prediction
  - one must know who/when/where people are infected or will become infected
  - Since diseases have an incubation period, there is a window of opportunity to save lives. Can also protect most susceptible population with prophylaxis measures.
- Contemporary Spatial Analysis 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?

New York Hospital Admission Data 2007 Count/Location Histogram


Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image USDA Farm Service Agency
Image PA Department of Conservation and Natural Resources-PAMAP/USGS
40°50'17.80" N 73°53'11.38" W elev 41 m
©2010 Google
Eye alt 140.86 km

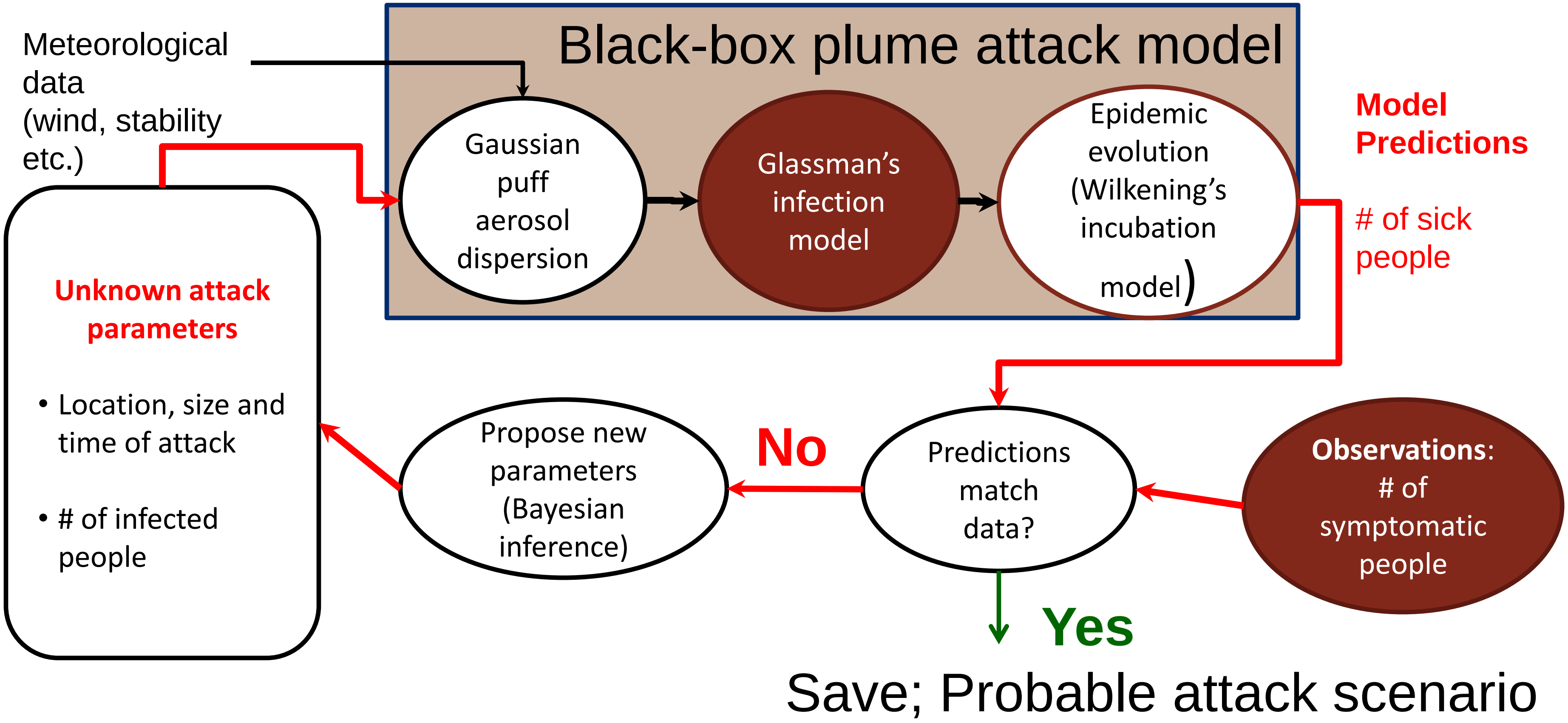
Plume Estimation Approach

- The key to forecasting infected people is to characterize the attack probabilistically
  - Location, size and time
  - Use a dispersion model + epidemic model to identify where the incubating and imminently susceptible people are (we already know the symptomatic ones)
- How? The model
  - Use a dispersion model to "spread" an aerosol and infect people with different doses
    - Inputs: location of release, amount of release
  - Use an epidemic model (say, for anthrax) to predict the evolution of the disease, given infected people with varying doses
    - Inputs: time of infection, # of infected people and their dosages.

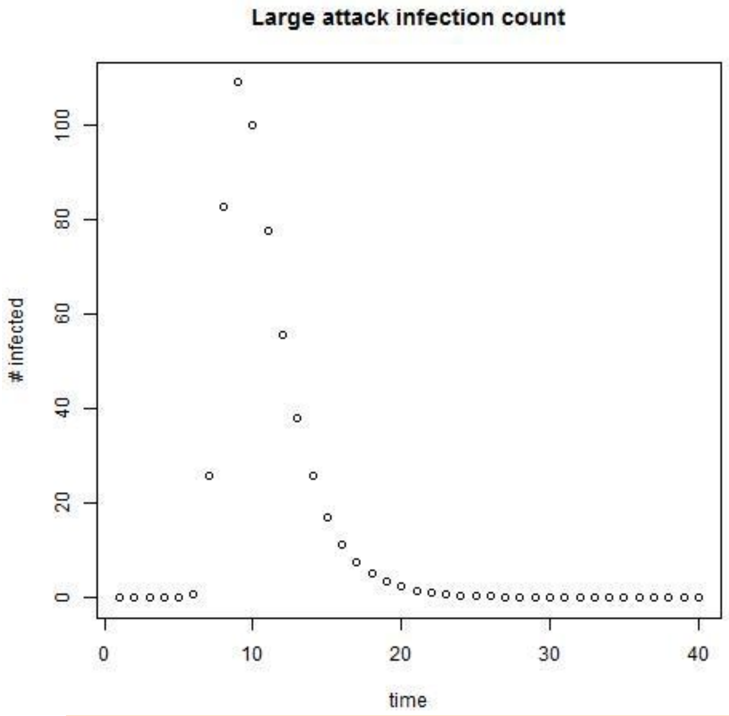
Plume Estimation Approach (cont.)

- 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
- 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?

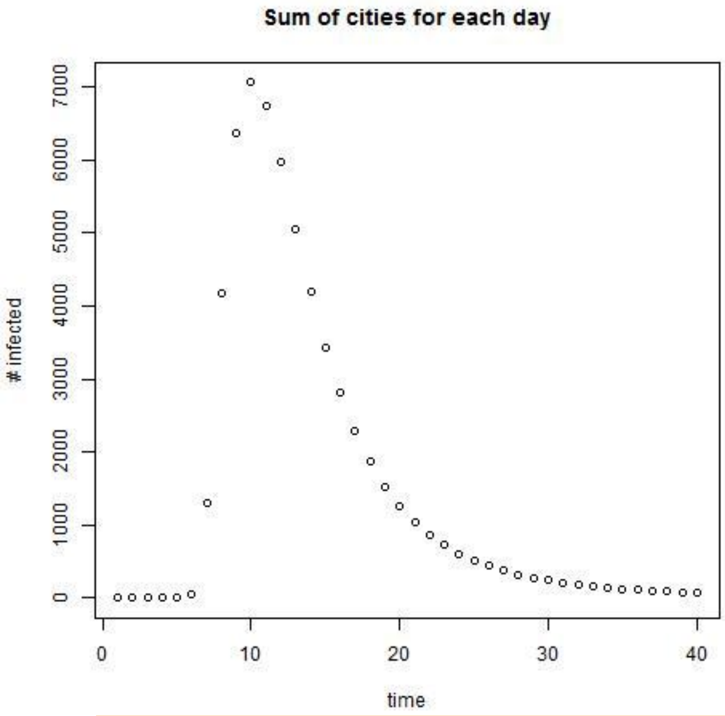
Inference – Fitting Models to Data: Plume Model


Meteorological data (wind, stability etc.)
Black-box plume attack model
Gaussian puff aerosol dispersion
Glassman's infection model
Epidemic evolution (Wilkening's incubation model)
Model Predictions
# of sick people
Unknown attack parameters
• Location, size and time of attack
• # of infected people
Propose new parameters (Bayesian inference)
No
Predictions match data?
Observations: # of symptomatic people
Yes
Save; Probable attack scenario

Case I –Attack with No Model Mismatch



Epidemic curve for a chosen zip-code

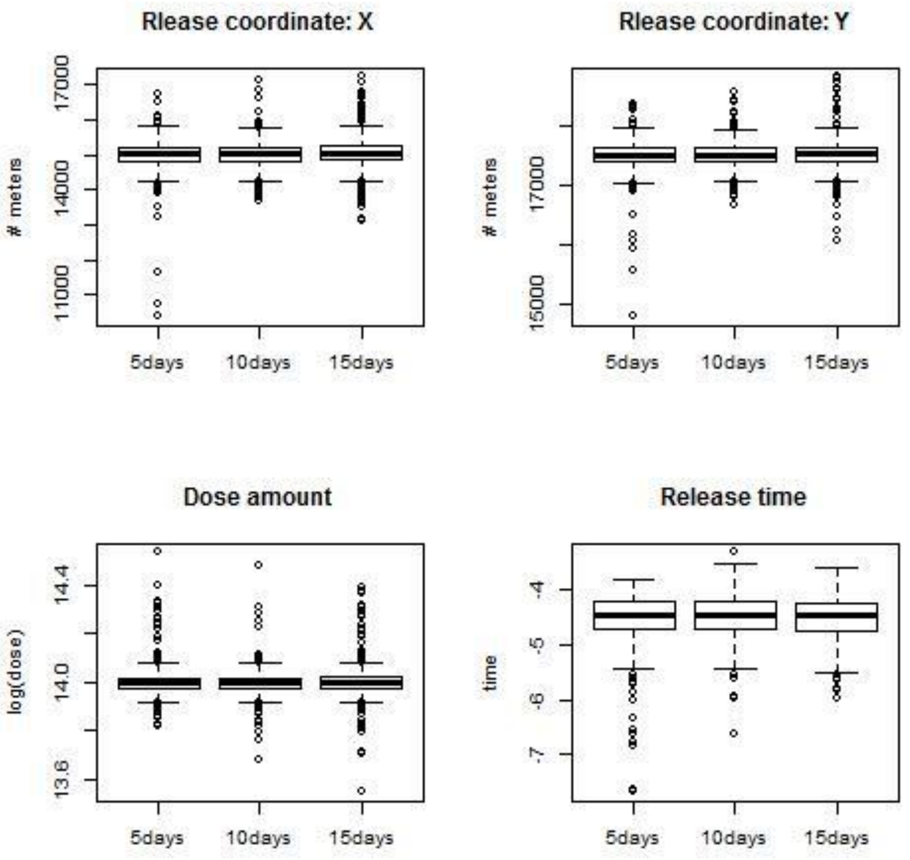


Epidemic curve for the entire city

- 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

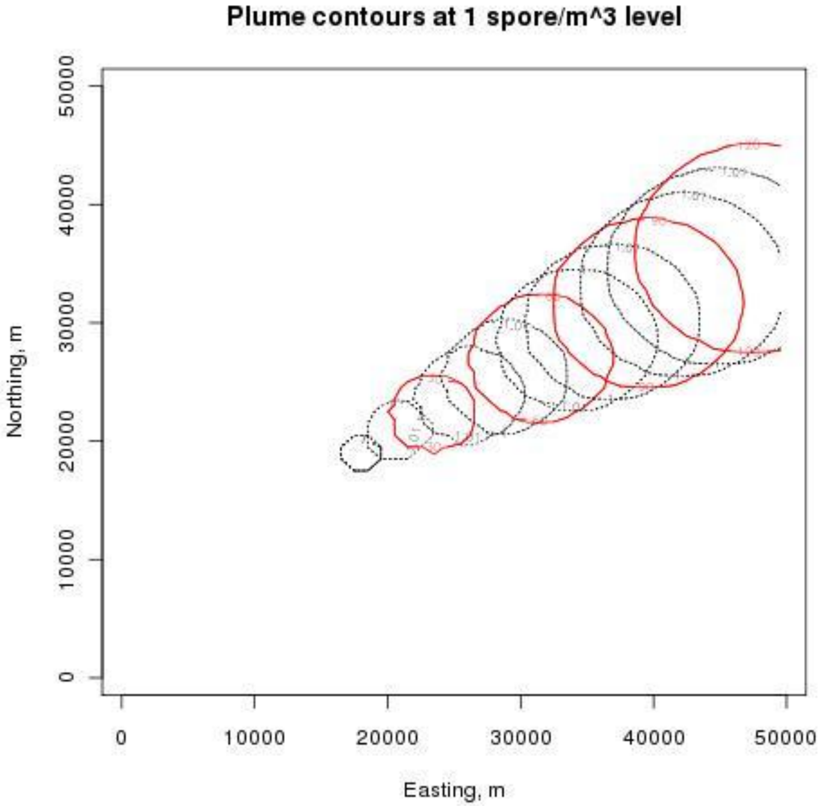
Inferred Location, Quantity and Time of Release

- Even 5 days of data is good enough
- True values:
  - X : 15,000 m
  - Y : 17,500 m
  - $\text{Log}_{10}(\text{Dose}) = 14$
  - Time = -5 days

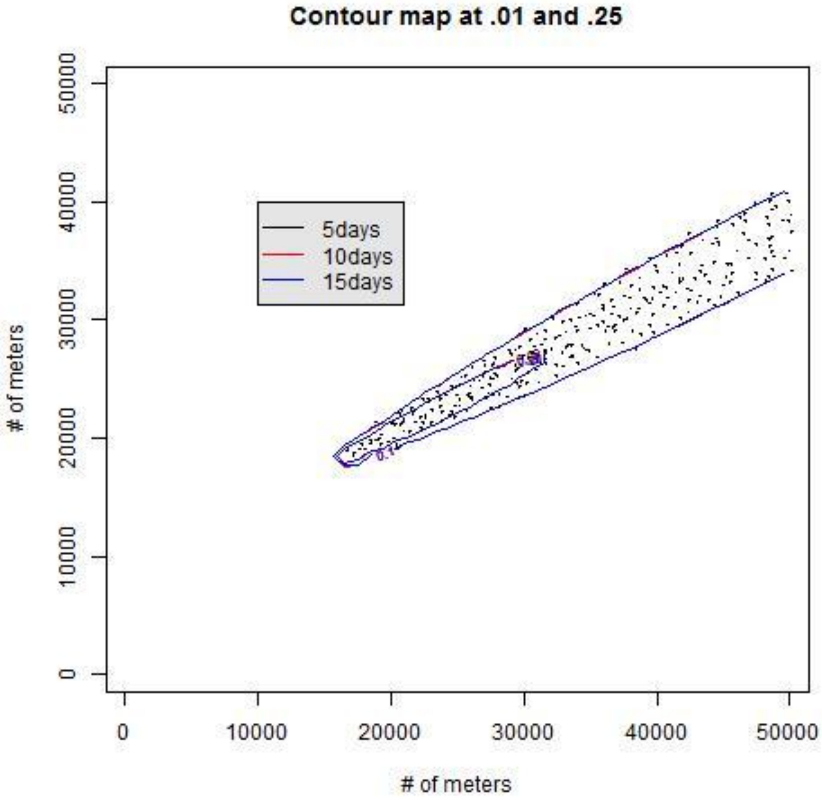


Inferred values of release location (X, Y), release size ($\log_{10}(Q)$) and release time. True values [15,000; 17,500; 14, -5]

Clusters – Observed and Predicted



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

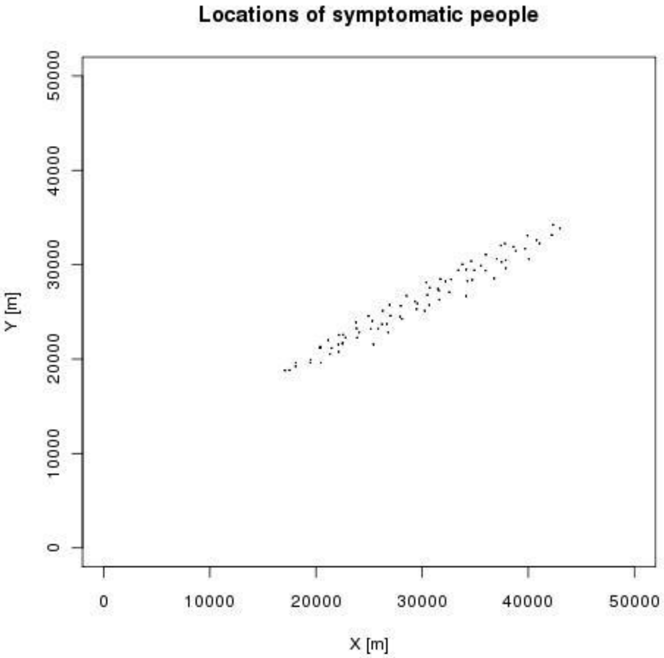


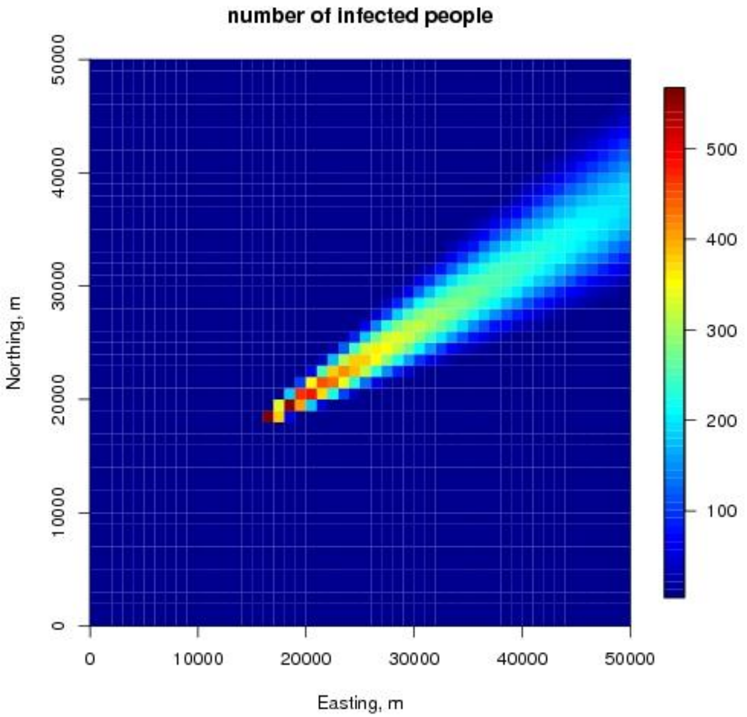
Contours show regions where 1% (outer) and 25% (inner) of the population are infected as a result of the release. Dots are individuals reporting.

Estimated Distribution of Infected People

- Spatial dissemination over a distributed population
- Estimate affected area from sparse (early) data
- Data = # of sick people / day / zip code

Distribution of symptomatic people on Day 5



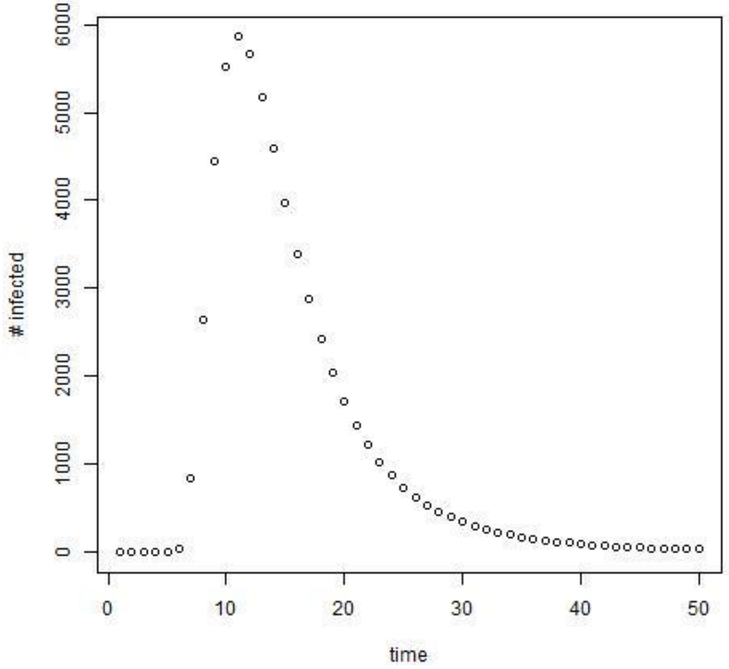


Estimated/true distribution of infected people

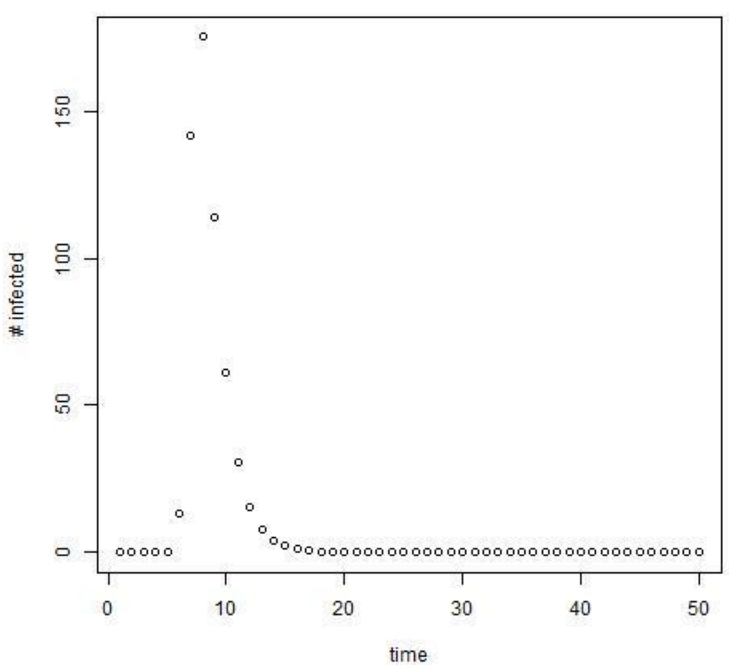
Naïve cluster analysis of the observations gives a wrong impression of true spatial distribution

Case II – Inference under Model Mismatch

Epidemic curve for the entire city



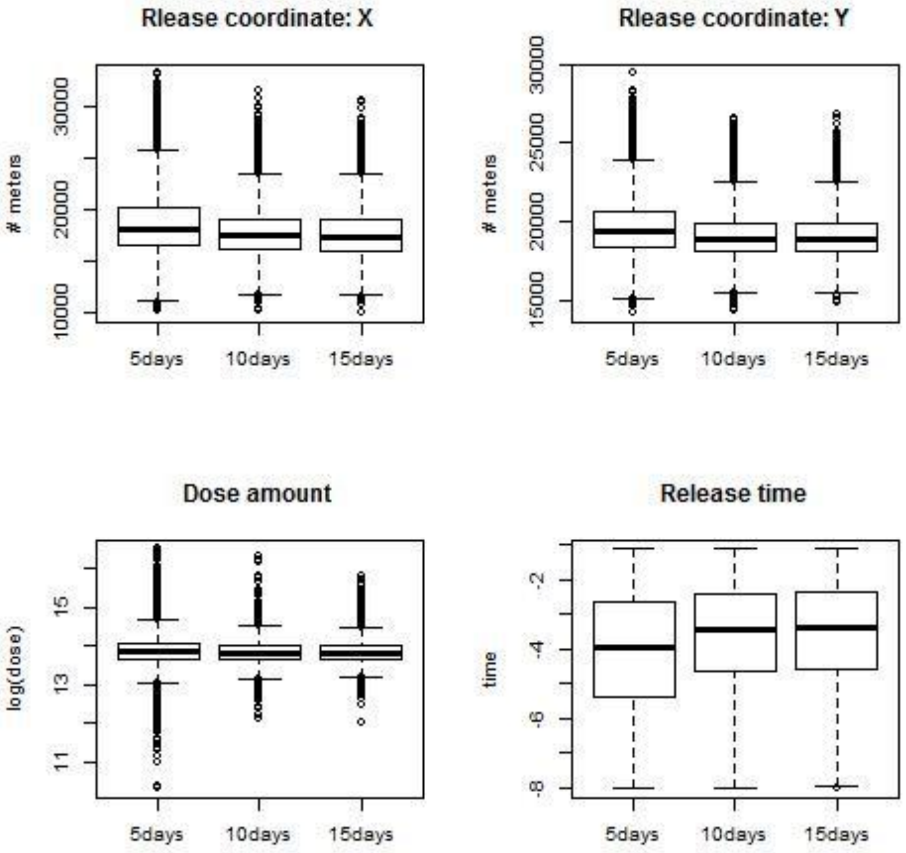
Epidemic curve for a chosen zip-code



- 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

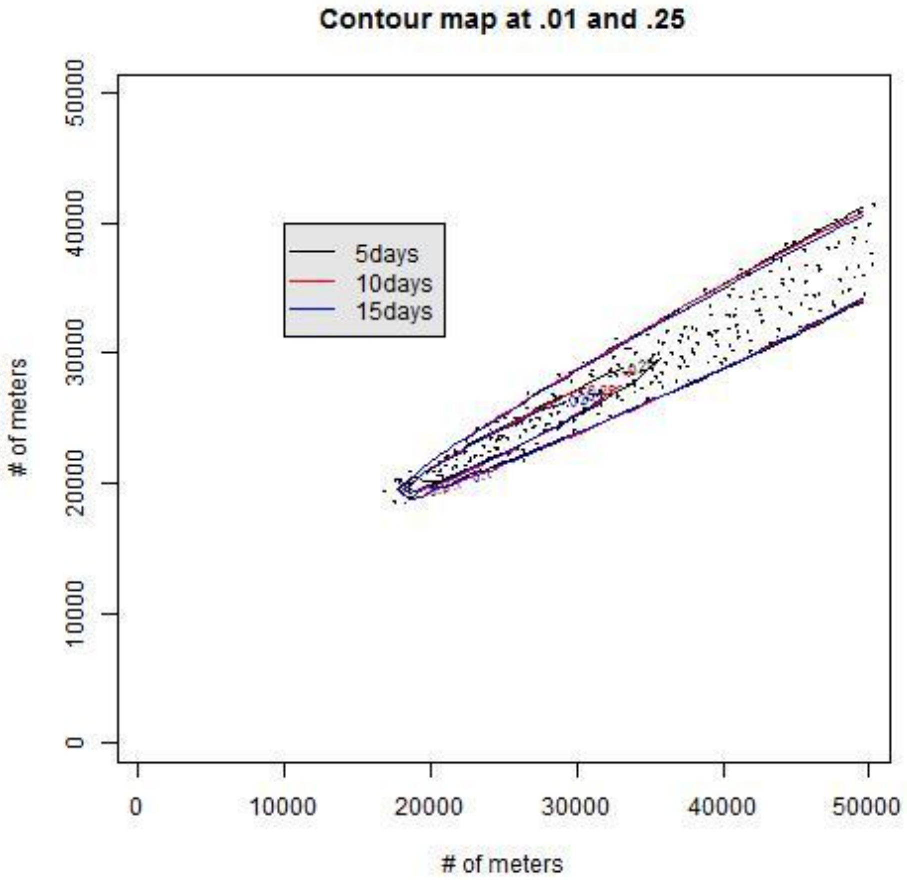
- Locations inferred wrongly – but by about 2 grid-cells (2 km)
- Underestimated release quantity
- Bigger uncertainties in time
- No improvement with addition of data (beyond 5 days)



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

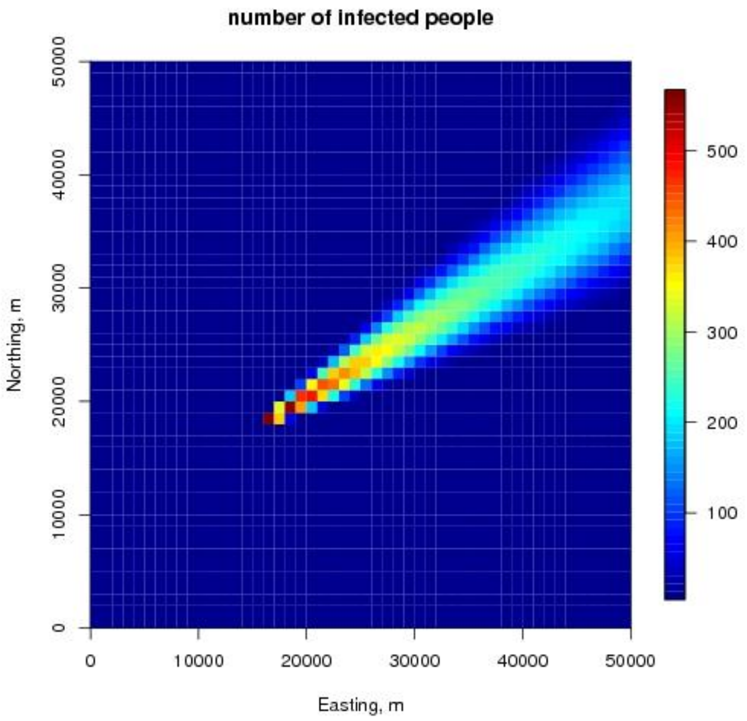
Contours – Observed and Predicted

Clustering still OK even with model mismatch

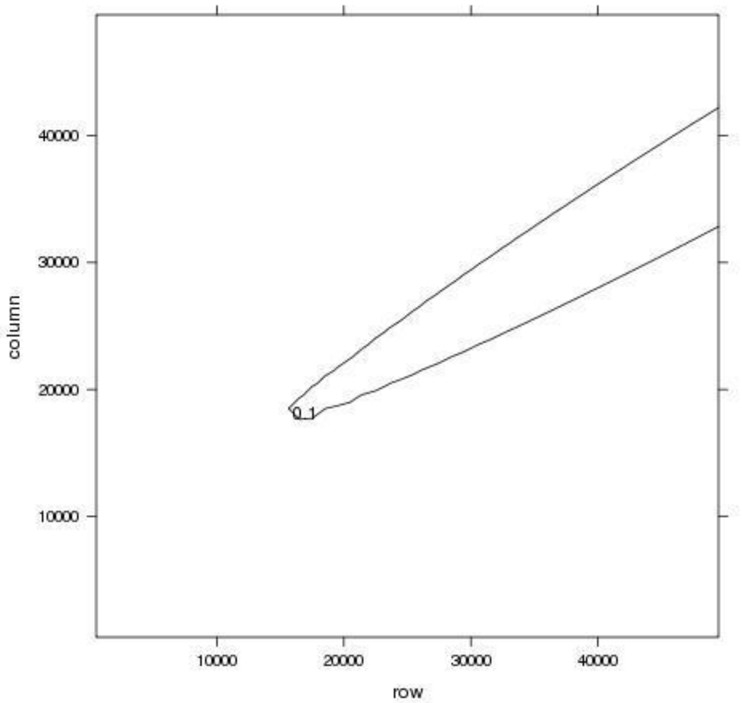


Contours show regions where 1% (outer) and 25% (inner) of the population are infected as a result of the release. Dots are individuals reporting.

Model-Informed Spatial Analysis



True distribution



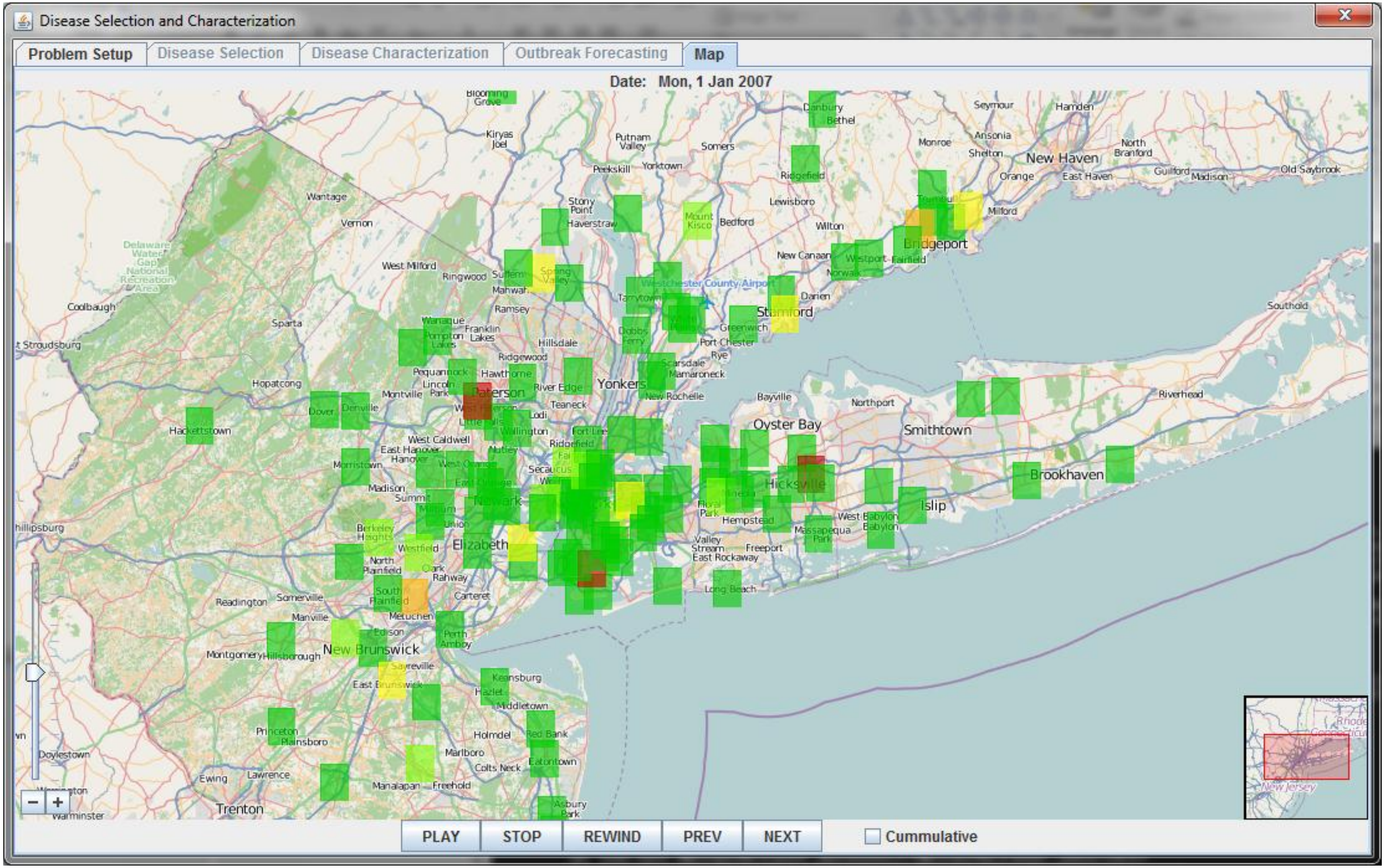
Reconstruction

Model-enabled reconstruction provides a better starting point for clustering/analyzing spatial biosurveillance data

Temporal-Spatio Visualization Prototype

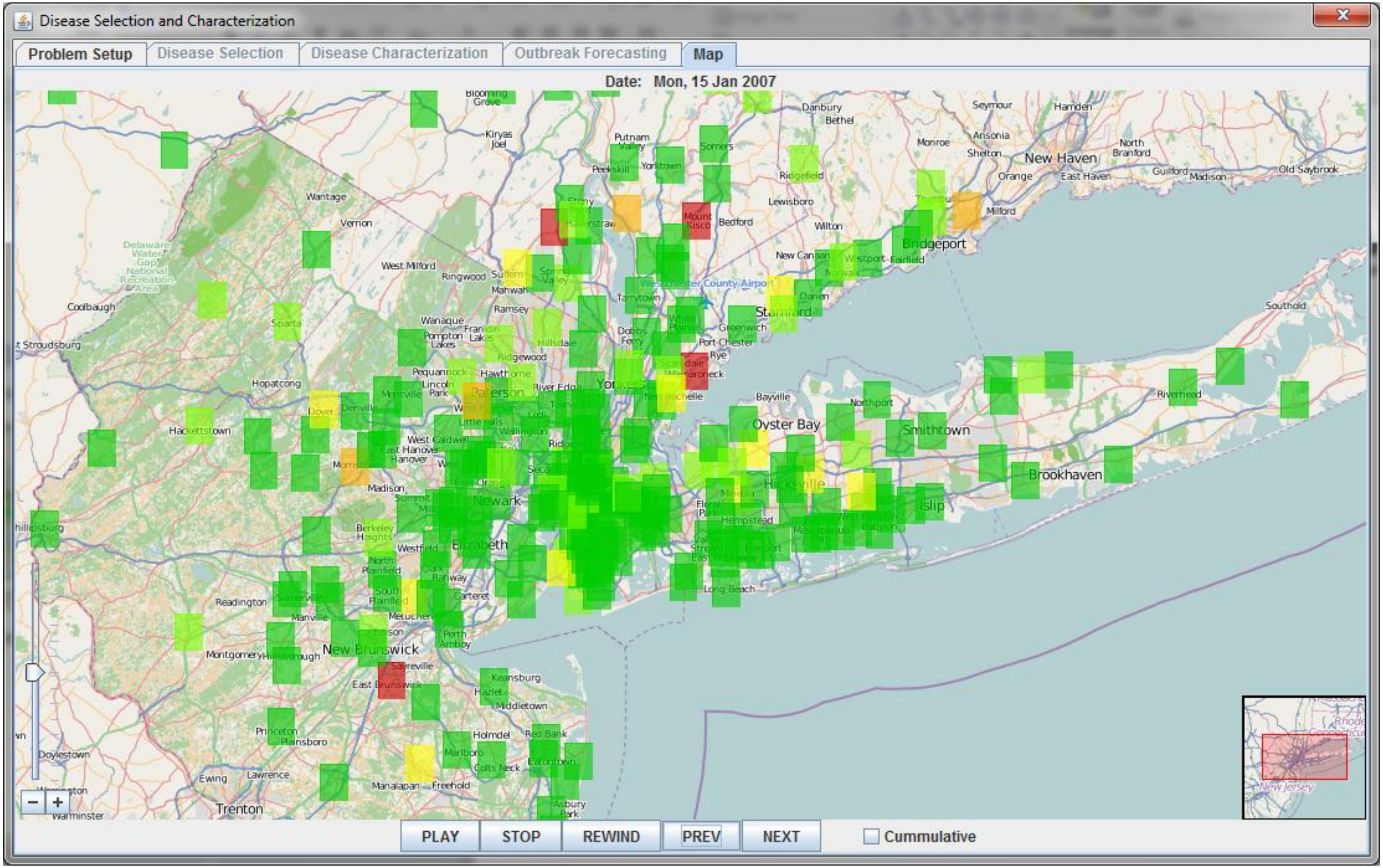
- Pure visualization alone is very useful for understanding outbreaks
- Prototype “Heat Map” of reports by zip code
  - Color based on number of events
  - Current day or cumulative counts
  - Animates changes in “playback” mode through time
- Future Enhancements Possible
  - Add source term estimation, etc.
  - Medical Resource Planning, etc.

Daily Report Heat Map



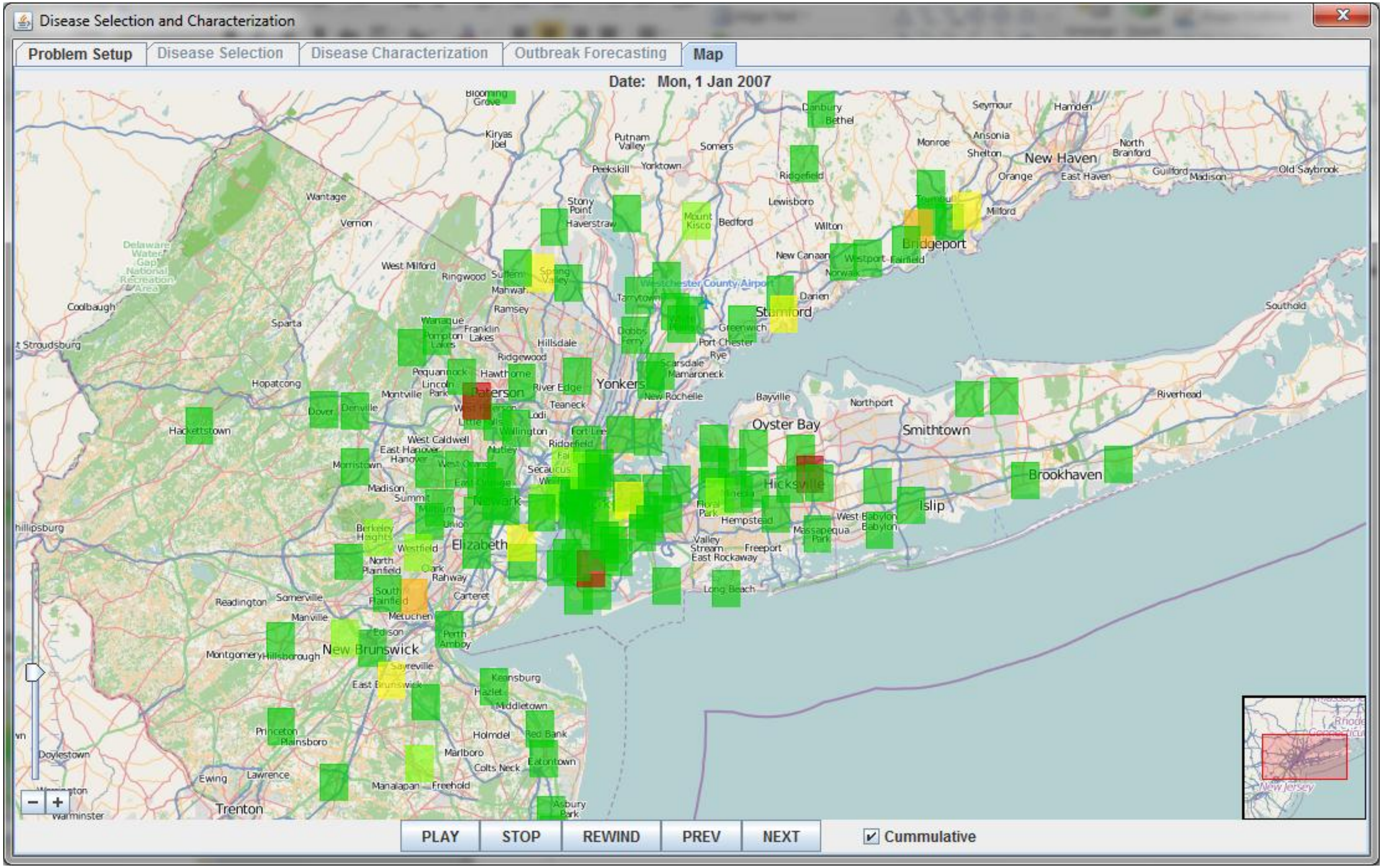
Date: Mon, 1 Jan 2007

Daily Report Heat Map



Date: Mon, 15 Jan 2007

Cumulative Report Heat Map

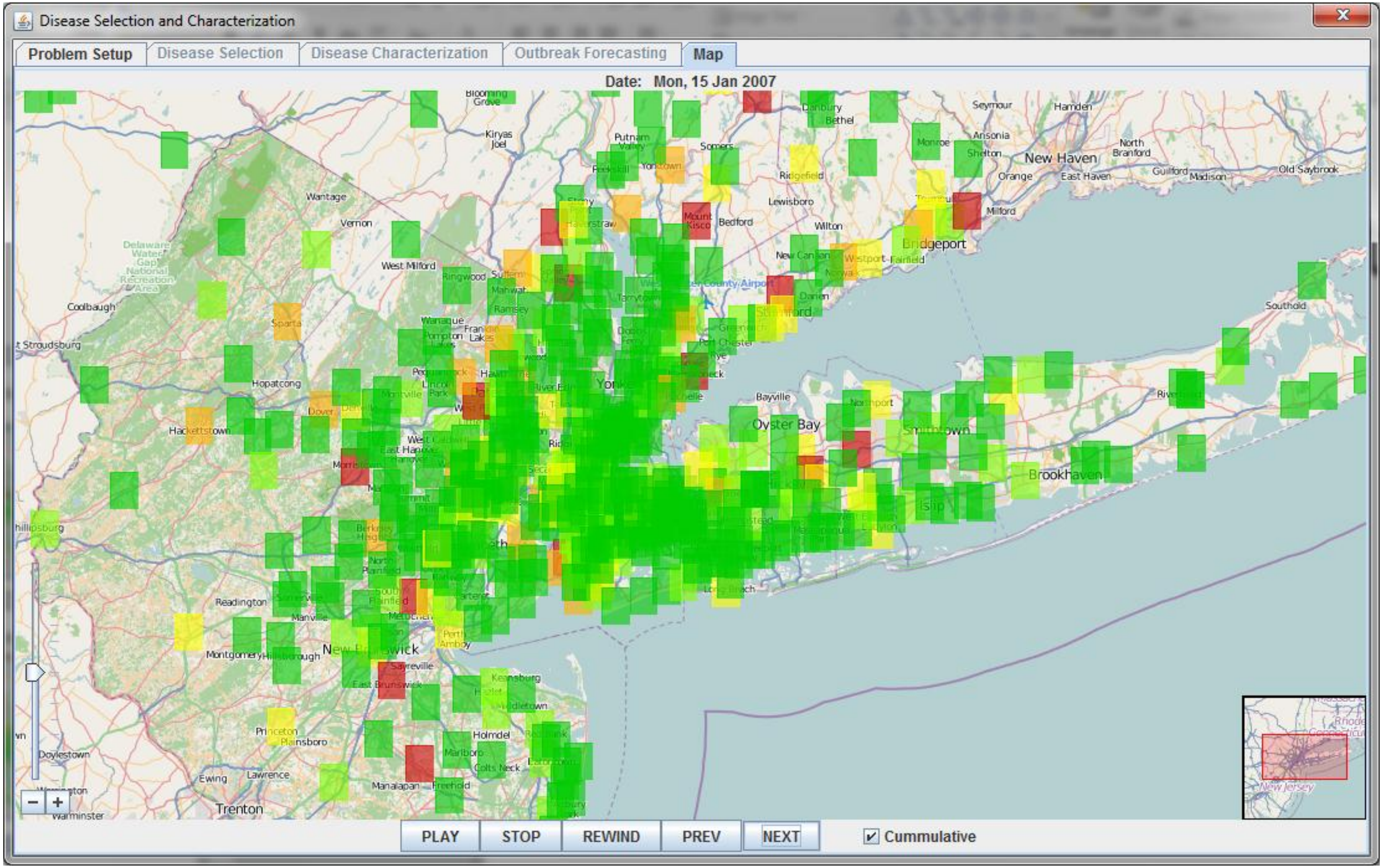


Disease Selection and Characterization

Date: Mon, 1 Jan 2007

☑ Cummulative

Cumulative Report Heat Map



Date: Mon, 15 Jan 2007

☑ Cummulative

Acknowledgements

This work is funded by the Defense Threat Reduction Agency (DTRA)
Ms. Nancy Nurthen at DTRA is the Program Manager.