Statistical Techniques for the Characterization of Partially Observed Epidemics

Jaideep Ray\textsuperscript{1}, Cosmin Safta\textsuperscript{1}, Karen Cheng\textsuperscript{2} and David Crary\textsuperscript{2}

\textsuperscript{1}Sandia National Laboratories, Livermore CA, \\
\textsuperscript{2}Applied Research Associates, Inc, Arlington, VA

\textbf{Acknowledgements:} The work was funded by DTRA, under contract HDTRA1-09-C-0034

Ms. Nancy Nurthen is the DTRA PM.  
Ms. Karen Cheng at ARA, is the Principal Investigator.  
Contact Info:  kcheng@ara.com and jairay@somnet.sandia.gov

SAND2010-7770C
Problem Statement

• **Aim:** To develop statistical techniques to characterize ongoing epidemics from partial biosurveillance data
  – Estimate # of index cases, time of infection, or infection rate
  – Do so with minimal data i.e., early in the outbreak
    • Data is a time-series of counts of ICD-9 codes
  – Quantify the confidence in the estimates

• **Motivation**
  – To provide initial conditions for disease models, to be used for planning medical interventions, resource allocation etc.
    • Disease models can be agent-based ones too
  – Can also be applied to historical epidemics, with case-counts as the data
    • Useful for obtaining disease model parameters for agent-based simulators.
Why Are Current Biosurveillance Methods Inapplicable?

- Current *biosurveillance* methods focus on detection
  - Based on anomaly detection
  - No model of the background
  - Or filtered out and this “disturbs” the detection

  “7 day moving average filters suppress exactly the short scale features that were the intended object of study”

- Current *characterization* methods for epidemics are used retrospectively
  - The epidemics are *fully* observed, not *partially* observed
  - The identity of the disease is known
  - The data consists of counts of people who have been diagnosed with the disease
    - It is *not* biosurveillance data with all its confounding issues
Difficulties with Using Biosurveillance Data

- Biosurveillance data (ICD-9 counts, OTC sales etc) is complex
  - Weekly & seasonal cycles; non-stationary structure
  - Symptom, not diagnosis, data (for timeliness)
- Characterization of epidemics with biosurveillance data requires:
  - Ability to model the background/endemic morbidity in real time
  - Detect the start of the epidemic
  - Extract the epidemic from the data
    - By “subtracting” the background

ILI ICD-9 stream from Miami (background / endemic morbidity)
Technical Challenges

• The components of the procedure are:
  – Detection of an outbreak from time-series data
  – Extraction of the outbreak from the background
    • Data for detection and extraction are ICD-9 streams with both the background/endemic and outbreak signal
    – Characterization of the outbreak (index cases, infection rate …)
• Biosurveillance data is partial, so …
  – All estimates are uncertain, and
  – The uncertainties need to be quantified
• Figures of merit
  – Delay between infection and detection
  – Cleanliness of the separation of background and epidemic
  – Closeness of inferred and true nature of outbreak
Detection of the Outbreak

• Based on sequential data assimilation using a Kalman Filter (KF)
  – Uses a simple model for daily ICD-9 counts (case-count)
  – Case-count model contains
    • A daily mean level and a cyclic weekly term
    • A quadratic, fitted to 4-week window of daily levels, for one-step-ahead predictions
  – KF also produce a measure of uncertainty in model predictions
    • KF covariance matrix
• Results in a model for the background morbidity
• Detection strategy:
  – Predict one-day ahead using quadratic model
  – If observation is greater than threshold, alarm (2-3 Std. Dev.)
  – Else, assimilate observation to obtain new mean level
Example with Synthetic Data

- **Simulated anthrax outbreak**
  - Small atmospheric release over a spatially distributed population (3 Million people)
  - 1125 index cases, with a range of doses
  - Includes visit delay
- **Background data for Miami (ICD-9 for ILI)**
  - Anthrax outbreak injected in on Day 130
- **KF starts fitting background model from Day 0**
- **Question:** How good is the background model
  - i.e. how many days to detection?

![Diagram showing the timeline of an anthrax outbreak and the detection process](image-url)
Detection Performance

- Based on Kalman Filters
  - Starts on Day 0
  - Creates a model of endemic ILI disease
- Detection:
  - One-day-ahead model predictions
  - Compared with observations
  - Significant deviation indicates an anomaly – detection!
  - In this case, detection took 5 days
  - Incubation: 3-4 days

Start: 130   Alarm: 135
Extraction of the Epidemic

• The “background” model can be “frozen” on the day of alarm
  – A quadratic is fitted to mean levels to determine local slope for forward projection
  – Weekly cycles derived previous data
  – KF formalism used for forward projection

• Questions:
  – How close are the model predictions to observations?

• Test this without the injected outbreak.

• Caveat: Model predictions will degrade in time

• Predictions up to 2 weeks ahead look good
  • But can this be used to extract the epidemic?
Extraction of the Epidemic Cont.

- Plot the difference between observations and predictions by frozen background model
- Estimate of the anthrax outbreak
  - Pretty good for 15 days
- However, it is a partial estimate
  - Extends only to the number of days of observations
- Can the partial anthrax outbreak be used for characterizing the attack?

Day 0 is day of release
Day 5 is day of detection
Characterization of the Anthrax Epidemic

- **Characterization:**
  - Estimation of the number of index cases, time of release, an average dose, and some parameters of the visit-delay model

- **Hypothesis:**
  - An anthrax incubation period model + a visit delay model can reproduce the epidemic curve
    - The quantities of interest are all parameters/inputs into this epidemic model
  - So given a partial epidemic curve, fitting an anthrax model should reveal the necessary model parameters

- **Questions:**
  - How much data is needed to estimate these parameters?
    - i.e., is less than 15 days of (good, normal background extracted) data sufficient?
  - What is the level of uncertainty in parameter estimates, as a function of (quantity of) data?
Bayesian Techniques to Solve the Problem

- The estimation is posed as a Bayesian inverse problem
  - Predicated on the extracted outbreak 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
Estimates of the Number of Index Cases

- Estimates of the number of index cases (in red).
- True figure in blue

Number of index cases bounded in 7 days;
Bounded to 2250 people out of original population of 3 Million;
Accurate to 20% after 9 days, post detection.
Incubation period is 3-4 days so will not get earlier than that.
Application to a Communicable Disease

• The technique can be applied to a communicable disease
  – Need to estimate infection rate (along with “usual” parameters)

• Assumptions for communicable diseases model
  – The infection rate rises & then falls smoothly in time
  – Index cases are a small fraction of the total number of victims

• A lightweight model can be created and fitted to data
  – The model of epidemic evolution is statistical (not AB)
  – Is used with MCMC, as before
  – Allows inferences to be drawn as PDFs

• Demonstrate with synthetic data
  – Simulate a plague epidemic using an AB model
A Communicable Disease Example

• The simulated plague epidemic
  – Includes visit-delay
  – Incubation is NOT dose dependent
• 100 index cases
  – Epidemic dies out in 40 days
  – 1500 victims, total
• Aim:
  – Estimate the total size of the epidemic
  – Also, the infection rate curve
  – Compare with the “true” figures from the simulation

- Red points: People turning symptomatic, daily (observed)
- Blue line: people being infected, daily (unobservable)
Estimation of the Final Epidemic Size

- The estimate improves (shorter error bars) with time
- Easier for large outbreaks
Conclusions

• Techniques appear promising to construct and integrate automated detect-and-characterize technique for epidemics
  – Working off biosurveillance data
  – Provides information on the particular/ongoing outbreak

• Potential use – in crisis management and planning, resource allocation
  – Parameter estimation capability ideal for providing the input parameters into an agent-based model
    • Index Cases, Time of Infection, infection rate

• Non-communicable diseases are easier than communicable ones
  – Small anthrax can be characterized well with 7-10 days of data, post-detection; plague takes longer
  – Large attacks are very easy