Biosurveillance: modeling and prediction of disease outbreaks using time series (or spatio-temporal data)

Peter F. Craigmile

http://www.stat.osu.edu/~pfc/

Workshop: Spatially-varying SDEs, with application to the biological sciences
Mathematical Biosciences Institute (MBI)
The Ohio State University, Columbus, OH
7-10 Jul 2015
**Background**

- Well-known, as well as previously uncharacterized infections continue to (re)emerge around the globe.

- To avoid casualties from outbreaks of these infections and from the potential criminal uses of bioagents, **biosurveillance systems** are needed that have the capacity to identify such outbreaks accurately and rapidly.

- **Accuracy** and **timeliness** of biosurveillance systems rests on the ability to model the uncertainty, severity, and aberrancy of clinical symptoms as expressed through the data monitoring system that are likely to portend disease outbreaks.
Problems in biosurveillance

- Shmueli [2005] summarizes the problems that biosurveillance systems, in general, pose to traditional statistical monitoring:

1. Biosurveillance data may not be independent or stationary;

2. Non-traditional data are assumed to contain earlier signature of an outbreak but this signal is weaker compared to actual diagnosis data;

3. Since there are no data that contain certain outbreaks, outbreak patterns – particularly as they would manifest in non-traditional data streams – are unknown;

4. In the assessment of biosurveillance methods, biosurveillance data are assumed to have no outbreaks (which may not be true).
Time series methods

- Time series methods are commonly used to detect disease outbreak signatures (e.g., signals due to influenza outbreaks and anthrax attacks) from varying respiratory-related diagnostic or syndromic data sources.

- Typically involves two components:

  1. Using time series methods to model the baseline background distribution (the time series process that is assumed to contain no outbreak signatures),

  2. Detecting outbreak signatures by looking for departures from the baseline series.
**Midwest pediatric emergency departments (EDs) [Craigmile et al., 2007]**

- We consider of modeling and detection of respiratory-related outbreak signatures based on chest radiograph ordering patterns from a number of pediatric **emergency departments (EDs)** located in the Midwestern region of the United States.

- We also have ambient temperature records collected in each city, as a covariate.
  - Can use the temperature series as a surrogate measure of the annual influenza season.

- Also, a patient visit count series is included in the models to account for variations between-EDs (like ED sizes) and within-EDs (day-of-week, for example).

- Addressing the fact that the underlying process is neither “independent or stationary”, our interest is to model the underlying “respiratory-complaint background”, using the available covariates and significant temporal dependencies present in the data.
An example dataset

- Emergency department (ED) data and temperatures.
  - ED visit daily-counts and chest radiograph daily-counts from 5 children-hospitals EDs (Minneapolis/St. Paul, Milwaukee, Chicago, Akron and Columbus, Jan 2003–Sep 2004).
  - Daily temperature series (from Daily Temperature Archive at The University of Dayton).

- Strong seasonal effect:
  - Visits and chest-radiographs: maximum peak in the winter and minimum peak in the summer.
  - Strong negative correlation with the daily temperature series.
**An approximate time series model**

- Let \( \{R_{k,t}\} \) denote the number of chest radiographs for the ED in city \( k \) (\( k = 1, 2, \ldots, 5 \)) on day \( t \) and let

\[
R_{k,t} = \beta_{0,k} + \beta_{V,k}V_{k,t} + \beta_{T,k}T_{k,t} + X_{k,t}.
\]

\( \{V_{k,t}\} \) are the visit counts.

\( \{T_{k,t}\} \) is the smoothed time series of temperatures.

\( \{X_{k,t}\} \) zero mean stationary time series.
Questions

• Can you build time series models (maybe even multivariate models) that can be used to model the baseline number of chest radiographs?
  
  – The approximate time series model that Craigmile et al. [2007] used can be improved!

• Design a method to detect outbreaks – how well does your method work?
  
  – You will need to have a way to simulate outbreaks!

• If my dataset is too old, feel free to analyze something more recent!
  
  e.g., Google flu trends https://www.google.org/flutrends/intl/en_us/us/
  (look at Cities)

  Weekly U.S. Influenza Surveillance Report
  
  http://www.cdc.gov/flu/weekly/
  (examine multiple weeks!)
A stochastic model for an anthrax outbreak

• Craigmile et al. [2007] consider a simple stochastic model for an inhalational anthrax outbreak.

• Work based on the work of Buckeridge et al. [2004] and Brookmeyer et al. [2005].

• Model incorporates:

  1. A stochastic model of infection and progression of the disease.

  2. A simple model of health-care-utilization that, on a day-by-day-basis, tracks the behavior of each infected individual.
Stages of inhalational anthrax outbreak

1. **Agent dispersal** of the anthrax spores.

2. Once spores are inhaled by a subject, in the **incubation** stage spores either **germinate** or are cleared out the lung.

3. After germination of the spores, the later stages of the disease are the **prodromal** and **fulminant** stages, followed by **death**.

- Brookmeyer et al. [2005] use a mathematical model to describe incubation of anthrax, based on the notion of the **attack rate** \( \Pr(\text{at least one of } D \text{ anthrax spores germinates}) \).

- Buckeridge et al. [2004] use a Gaussian plume model for the dispersion of anthrax spores, and then use lognormal distributions to model the incubation, prodromal and fulminant duration times.

- **We assume an outbreak that affects a fixed number of children, \( N \), in an enclosed area** (with no spread of the spores outside of this area), and no second infection.
A simplified health-utilization model

1. During the **incubation** stage, we assume no people enter the ED.

2. During the **prodromal** and **fulminant** stages we assume a simple **Markov model** of utilization:

   Each infected subject is a Bernoulli event, independent across days, with the probability of entering the ED on a weekday being $P_d$ and on a weekend, $P_w$.

3. At end of the **fulminant** stage we assume the subject exits the system (i.e., dies).

- Buckeridge et al. [2004] use a similar model, but do not report their probabilities.

- This model could easily be extended to include, e.g., varying probabilities of entering the ED by stage, and/or other ways to exit the system such as a pharmacy visit.
The simulations presented in Craigmile et al. [2007]

For each city:

- We fit the regression time series model using the first year of data (training data).
- We evaluate our method for detection of outbreaks in the second year (test data).
- One hundred times:

  1. We simulate an outbreak with attack rate 0.5, starting on April 1st.
  2. ED utilization probabilities: $P_d = 0.25$, $P_w = 0.4$.
  3. We calculate the proportion of true non-detects before the outbreak (the specificity), and the proportion of true detects during the outbreak (the sensitivity), for four different detection algorithms.
References


