




# Anomaly Detection Techniques for Biosurveillance Applications

Karen Cheng and David Crary  
Health Effects And Medical Response Group  
Applied Research Associates, Inc., Arlington, Virginia, USA

Jaideep Ray and Cosmin Safta  
Sandia National Laboratories  
Livermore, California, USA


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

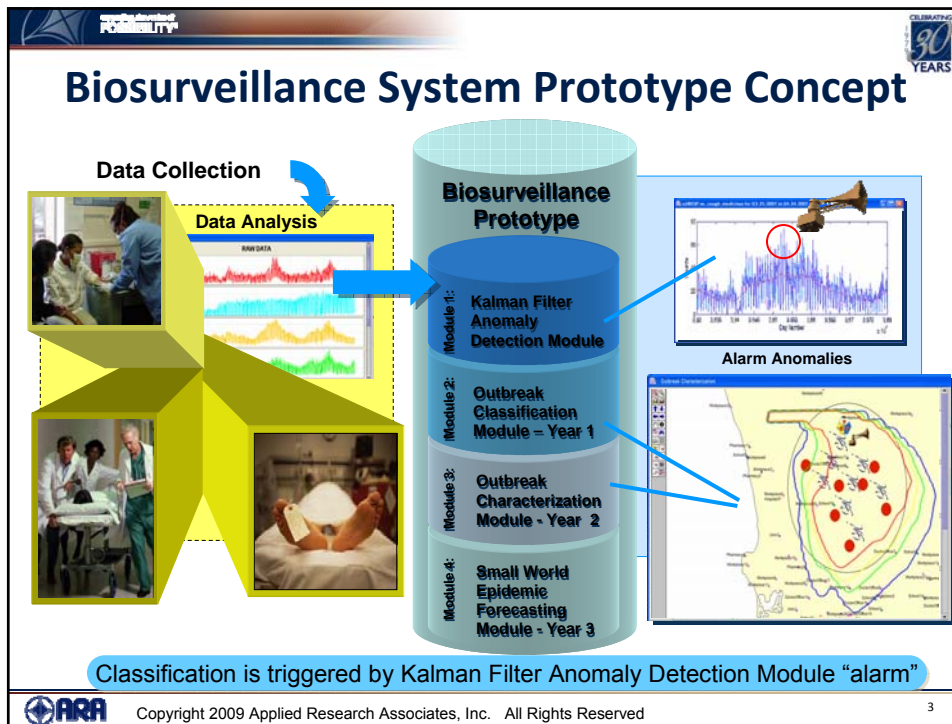
 Copyright 2009 Applied Research Associates, Inc. All Rights Reserved 1

## Context

- Anomaly detection techniques described in this talk are part of a larger, ongoing project to detect and classify diseases using Bayesian techniques on the basis of partially observed epidemic data (further discussed by C. Safta, *et al.*, this conference).
- For this study, observed data is a time series of morbidity counts (derived from ICD-9 codes, etc.).
- Bayesian techniques used in fitting time series data to epidemic models require an accurate 'starting date' for calculations, when the data become statistically distinguishable from noise.
  - Bayesian techniques are computationally expensive

 Copyright 2009 Applied Research Associates, Inc. All Rights Reserved 2



- Anomaly Detection Approach**
- Anomaly detection: based on well-founded structural time series models within a Kalman Filter framework to provide a statistical characterization of anomalies in time series data
  - The technique applies to any informative time series data including both traditional and non-traditional sources (morbidity data, veterinary data, drug requisition)
    - Applicable to other domains
  - Provides one step ahead prediction capabilities and automatic error propagation
  - Easily modified to incorporate weekly variations
  - Naturally models slope and trends in the data without using ad-hoc filtering techniques (Bloom, Buckeridge, and Cheng, "Finding leading indicators for disease outbreaks: Filtering, cross-correlation, and caveats", JAMIA, 14, 2007)
- Copyright 2009 Applied Research Associates, Inc. All Rights Reserved 4



## Advantages of the Kalman Filter Approach to Anomaly Detection

- Structural Time Series provide a method for specifying a model of an underlying dynamical mechanism producing the time series
  - Models can directly incorporate knowledge of time series dynamics
  - Traditional modeling techniques such as Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) models are parameterizations of the second order properties of time series – the question of the underlying dynamics is not addressed
- Can easily handle missing values
  - Essential for modeling medical time series
  - Traditional ARIMA techniques for handling trends (or other non-stationary) use differencing, which propagates missing values
- Forecasting handled in a formal statistical framework
  - Traditional forecasting techniques (exponential smoothing, Holt-Winters forecasting) do not provide forecast error estimates



## Testing the Method

- To test our techniques, we have created time series data from simulated anthrax outbreaks superimposed on background morbidity data
  - Background morbidity was simulated using Gaussian noise superimposed on a deterministic 'seasonal' component
- Anthrax incubation period derived from dose dependent Wilkening A2 model. (D.A. Wilkening, PNAS 103(20): 7589-7594, May 2006)
- Delay to "Present for Care" is governed by a lognormal distribution (reference W.R. Hogan, G.L. Wallstrom:  
<http://www.galaxy.gmu.edu/QMDNS2007/QMDNS2007-booklet.pdf>)



**Study Case Scenario**

**Anthrax attack**

- 22375 infected with a spatially variable dose
- Average dose: 2748 spores
- 1<sup>st</sup> symptoms [24, 48] hrs after infection
- Superimposed on background morbidity on Day 50

Copyright 2009 Applied Research Associates, Inc. All Rights Reserved 7

**Surveillance Methods**

- Outbreak monitored using one-step ahead forecasts from a Kalman filter based model assuming random walk in linear trend and local level

○ KF Prediction

— Observations

Epidemic starts on Day 50

- Indistinguishable from background
- Detection on day 56

- One-step ahead projection of time series value includes an error estimate, which can be used to set an anomaly detection threshold for incoming data on a standardized residual
- Incoming data points that exceed the threshold are used as the input to the Bayesian Classification model

Copyright 2009 Applied Research Associates, Inc. All Rights Reserved 8



**Statistical Characterization of Kalman Filter Performance**

- Tests were performed with 500 instances (background noise varies) of this scenario to check statistical properties:
  - False alarm rate: ~3% (as expected from detection threshold)
  - Average detection time 4 days +/- 1 day
  - Missed detections: 0
  - Q-Q plots show residuals are distributed normally

Technique provides a good method for eliminating non-stationarity in data, with low number of missed detections

**Residual Q-Q Plot**

Copyright 2009 Applied Research Associates, Inc. All Rights Reserved

11

**Conclusions and Further Work**

- Preliminary results indicate that Kalman Filter surveillance techniques are a useful technology for providing early warning of impending epidemics
  - Accurately models seasonal effects
  - Provides timely indication of anomalies essential for Bayesian disease characterization study (Safta *et al.*, this conference)
- Further work
  - Modify model to handle weekly cycles
  - Test on real data
    - We have ICD-9 codes corresponding to flu symptoms for Miami

Copyright 2009 Applied Research Associates, Inc. All Rights Reserved

12

## Acknowledgements

This work is funded by DTRA Contract HDTRA1-09-C-0034.

Dr. Christopher Kiley of DTRA is the S&T Manager  
Christopher.Kiley@dtra.mil