Prospective Biosurveillance for Early Detection of Disease Outbreaks

Authors:

- Mahmood Akhtar[#]
- Blanca Gallego
- Andy Yi-Chih Shiue
- Vitali Sintchenko

[#]Now with 'The University of Sydney'
email: m.akhtar@usyd.edu.au

Outline

- Background
  - Scan Statistic
  - Prospective Disease Surveillance
  - Elements of the Problem
  - Kulldorff’s SSS
  - Bayesian SSS
- Simulation Results
- Summary

Background

- **Scan Statistic**
  - To detect a local excess of events
  - Naus JI, (1965)
  - Main idea:
    - $[a, b]$: win$[t, t + w] \rightarrow w < b - a$
    - For all 't': record the maximum number of events in the window, and compare to its distribution under the null hypothesis of a purely random Poisson Process
  - Scope: detect disease clusters, use in brain imaging, astronomy, etc

- **Prospective Disease Surveillance**
  - Objective: detect spatial clusters of disease cases resulting from disease outbreak
  - Surveillance on daily basis, with the goal of finding emerging epidemics as quickly as possible.
  - Given data: no. of cases, spatial locations,
  - Rely on related observable quantities such as no. Of ED visits or OTC drug sales

- **Elements of the Problem**
  - Daily data collected for a set of spatial locations s_i
  - At each s_i, we have a count c_i, and an underlying baseline b_i
  - Goal: to find if there is any spatial region S (set of locations s_i) for which counts are significantly higher than expected, given the baseline
  - The set of all regions S in grid G is searched

- **Cluster Detection: two main goals**
    - To pinpoint the location, shape and size of each potential cluster
    - To determine (test) if a potential cluster is likely to be a "true" cluster or chance occurrence

"We compare the null hypothesis 'H_0' of no clusters against some set of alternative hypotheses '$H_1(S)$', each representing a cluster in some region or regions 'S' subset of 'G'

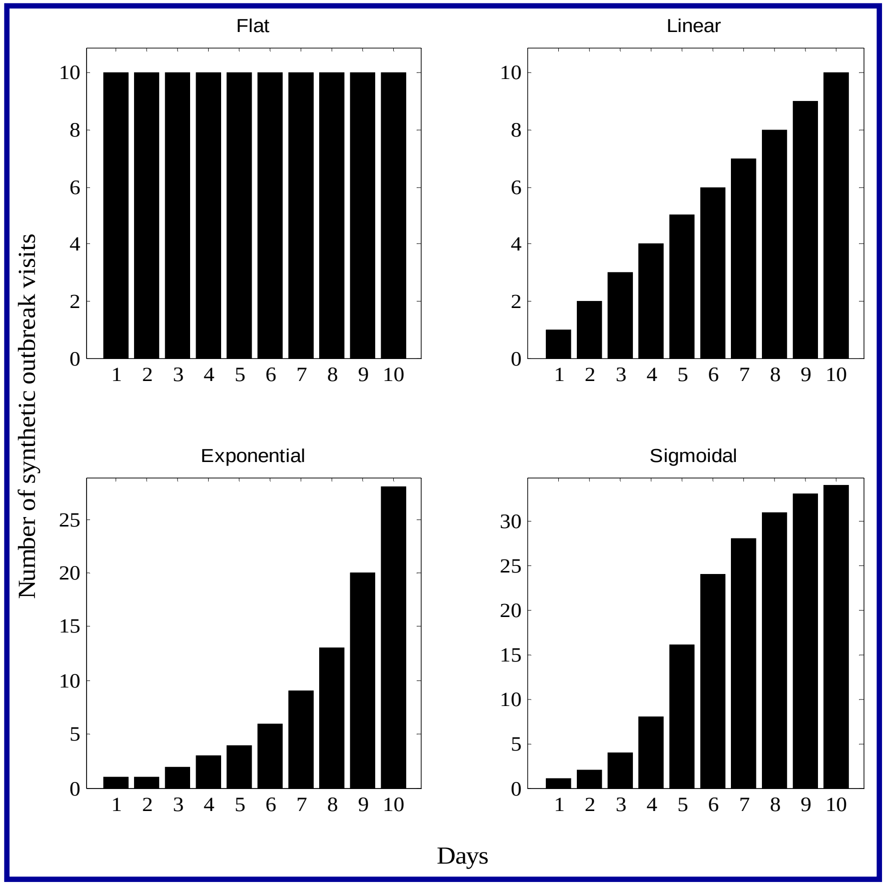
- **Kulldorff's SSS** **(M Kulldorff, 1997)**
  - One of the most important statistical tool for cluster detection
  - Searches over a given set of spatial regions, finding those regions which maximize a LR statistic
  - Statistical significance determined through randomization testing, very time consuming, computationally infeasible for large datasets
  - Other issues: no use of prior information, highly dependent on the MLE

- **Bayesian SSS (DB Neil et al., 2006)**
  - Uses prior information about size and shape
  - More flexible, less prone to overfitting
  - Increased power to detect clusters and much faster runtime (randomization testing is no more required)
  - Testing via posterior probabilities of each potential cluster
  - Complexity of $O(N^4)$ vs. $O(RN^4)$

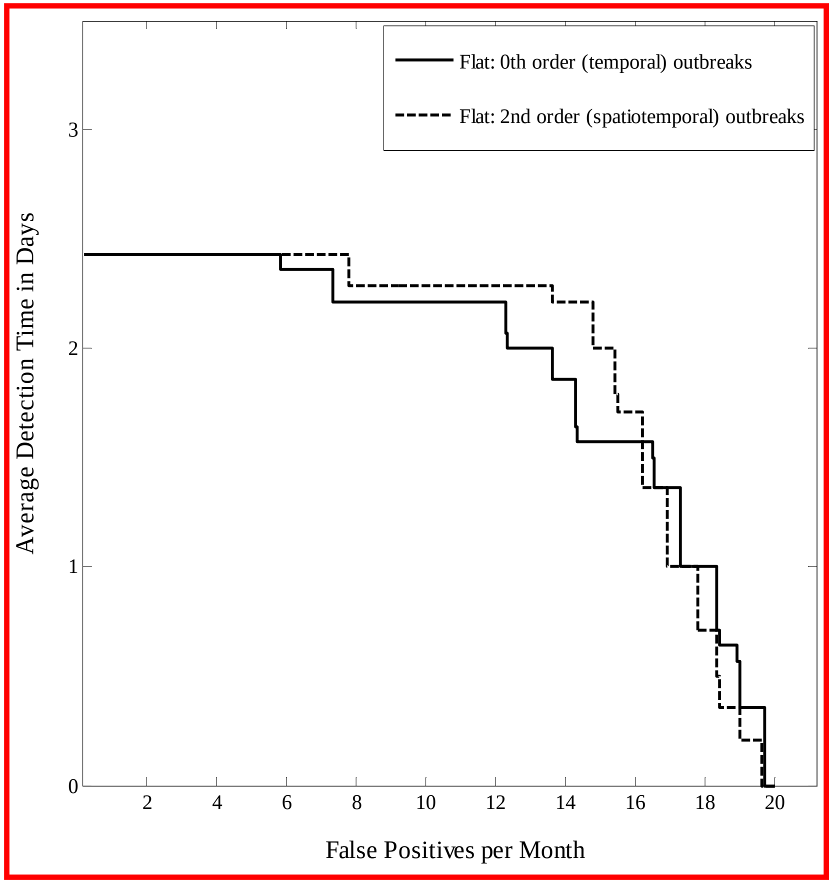
Simulation Results

- **Methods**
  - Bayesian SSS: implemented in Java
  - Simple exact algorithm (D Agarwal et al., 2006): underlying spatial scan for Bayesian model
- **Datasets**
  - Spatial locations: 79 postcodes of Sydney
  - Real Salmonella outbreaks: training set
  - Simulated spatiotemporal outbreaks: testing set, generated using Matlab / SAS packages
  - Evaluation measure: AMOC (Fawcett & Provost, 1999)

Simulated Outbreaks

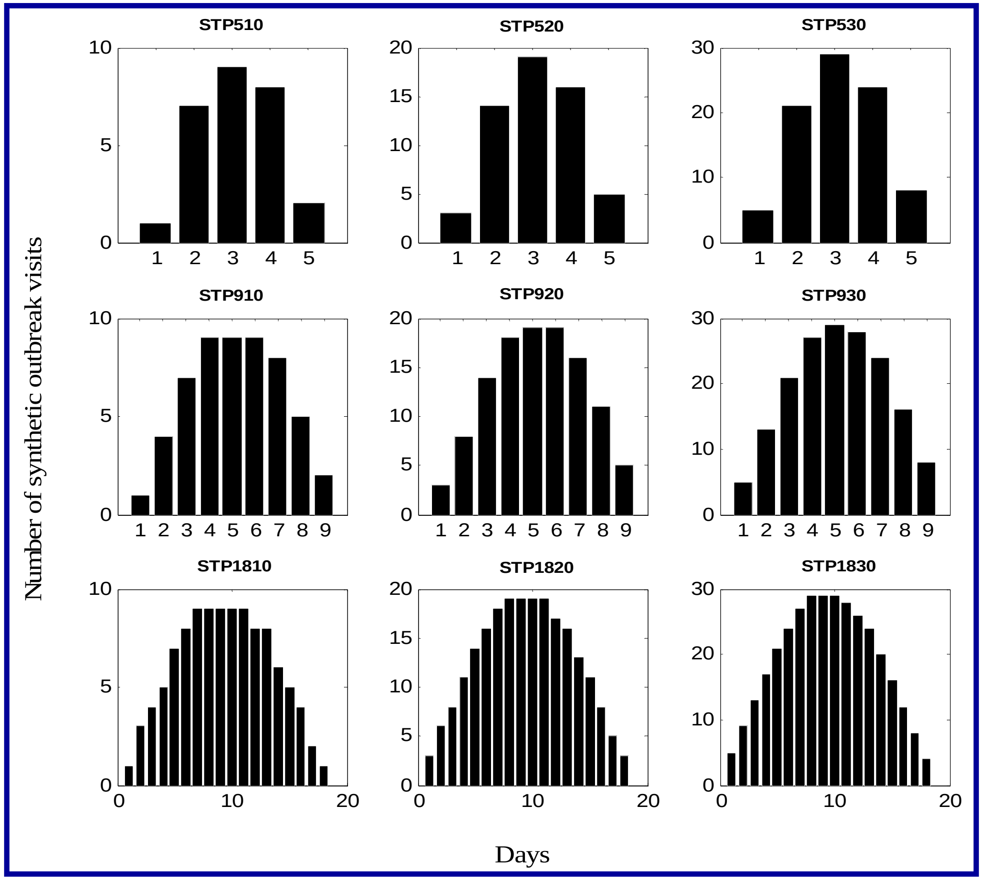


AMOC Analysis

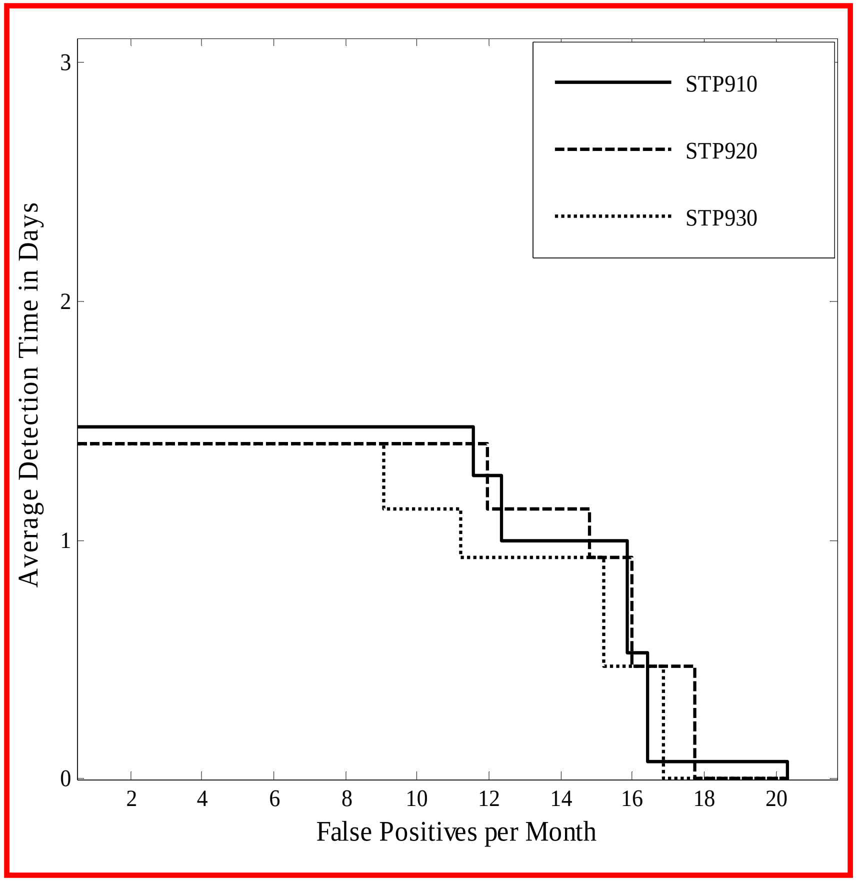


UNSW Centre for Health Informatics
The University of Sydney
brain&mind research institute

Simulated Outbreaks



AMOC Analysis



Summary

- Bayesian SSS model was implemented for prospective biosurveillance
- True outbreaks were used for estimation of model parameters
- Simulated outbreaks were used for performance measurement using AMOC curves
- Overall, the accuracy and timeliness results of this initial evaluation are encouraging
- Further testing on more diverse simulated and real outbreaks is required