2008-08

Improving Biosurveillance: Protecting People as Critical Infrastructure

Fricker, Ronald D.

Invited speaker, "Improving Biosurveillance: Protecting People as Critical Infrastructure, CNO DFP Critical Infrastructure Vulnerability Workshop, August 2008.
Improving Biosurveillance: Protecting People as Critical Infrastructure

Ronald D. Fricker, Jr.
August 14, 2008
What is Biosurveillance?

  - “The term ‘biosurveillance’ means the process of active data-gathering ... of biosphere data ... in order to achieve early warning of health threats, early detection of health events, and overall situational awareness of disease activity.” [1]
  - “The Secretary of Health and Human Services shall establish an operational national epidemiologic surveillance system for human health...” [1]

- Epidemiologic surveillance:
  - “…surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response.” [2]

Can Think of It Like a Large System of Sensors

- **Issue:** False alarms a serious problem
  - “…most health monitors… learned to ignore alarms triggered by their system. This is due to the excessive false alarm rate that is typical of most systems - there is nearly an alarm every day!” [1]

The Problem in Summary

• Goal: Early detection of disease outbreak and/or bioterrorism

• Issue: Currently detection thresholds set naively
  – Equally for all sensors
  – Ignores differential probability of attack

• Result:
  – High false alarm rates
  – Loss of credibility


“…most health monitors… learned to ignore alarms triggered by their system. This is due to the excessive false alarm rate that is typical of most systems - there is nearly an alarm every day!”[1]
• Each hospital sends data to CDC daily
  – Let $X_{it}$ denote data from hospital $i$ on day $t$
  – If no attack anywhere $X_{it} \sim F_0$ for all $i$ and $t$
  – If attack occurs on day $\tau$, $X_{it} \sim F_1$, $t = \tau, \tau+1, \ldots$
    • Assume only one location attacked
• Threshold detection: Signal on day $t^*$ if
  \[ X_{it^*} \geq h_i \]
  for one or more hospitals
• Each hospital location has an estimated probability of attack:
  \[ p_1, \ldots, p_n, \sum_i p_i = 1 \]
Idea of Threshold Detection

Distribution of Background Disease Incidence ($f_0$)

Distribution of Background Incidence and Attack/Outbreak ($f_1$)

Probability of a true signal:
$$\int_{x=h}^{\infty} f_1(x)dx = 1 - F_1(h)$$

Probability of a false signal:
$$\int_{x=h}^{\infty} f_0(x)dx = 1 - F_0(h)$$
It’s All About Choosing Thresholds

- For each hospital, choice of $h$ is compromise between probability of true and false signals

**ROC Curve**

| $\Pr(\text{signal | attack})$ |
|-----------------------------|
| 1.0                         |
| 0.8                         |
| 0.6                         |
| 0.4                         |
| 0.2                         |
| 0.0                         |

| $\Pr(\text{signal | no attack})$ |
|-----------------------------|
| 0.0                         |
| 0.2                         |
| 0.4                         |
| 0.6                         |
| 0.8                         |
| 1.0                         |

- No Attack/Outbreak Distribution
- Attack/Outbreak Distribution
• It’s simple to write out:

\[
\Pr(\text{detection}) = \sum_i \Pr(\text{signal|attack}) \Pr(\text{attack})
\]

\[
\mathbb{E}(\text{# false signals}) = \sum_i \Pr(\text{signal|no attack})
\]

• Express it as an optimization problem:

\[
\max_h \sum_i \left[ 1 - F_1(h_i) \right] p_i
\]

\[
\text{s.t. } \sum_i \left[ 1 - F_0(h_i) \right] \leq \kappa
\]
Some Assumptions

- Hospitals are spatially independent
- Monitoring standardized residuals from model
  - Model accounts for (and removes) systematic effects in the data
  - Result: Reasonable to assume $F_0 = \mathcal{N}(0,1)$
- An attack will result in a 2-sigma increase in the mean of the residuals
  - Result: $F_1 = \mathcal{N}(2,1)$
- Then, problem is: 
  $$\min_h \sum_i \Phi(h_i - 2)p_i$$
  subject to: 
  $$\sum_i \Phi(h_i) > n - \kappa$$
### Ten Hospital Illustration

<table>
<thead>
<tr>
<th>Hospital $i$</th>
<th>$p_i$</th>
<th>Common Threshold #1</th>
<th>Optimal Threshold ($h_i$)</th>
<th>Common Threshold #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.797</td>
<td>2.189</td>
<td>1.068</td>
<td>1.310</td>
</tr>
<tr>
<td>2</td>
<td>0.064</td>
<td>2.189</td>
<td>3.602</td>
<td>1.310</td>
</tr>
<tr>
<td>3</td>
<td>0.056</td>
<td>2.189</td>
<td>3.732</td>
<td>1.310</td>
</tr>
<tr>
<td>4</td>
<td>0.048</td>
<td>2.189</td>
<td>3.915</td>
<td>1.310</td>
</tr>
<tr>
<td>5</td>
<td>0.013</td>
<td>2.189</td>
<td>4.656</td>
<td>1.310</td>
</tr>
<tr>
<td>6</td>
<td>0.006</td>
<td>2.189</td>
<td>4.736</td>
<td>1.310</td>
</tr>
<tr>
<td>7</td>
<td>0.006</td>
<td>2.189</td>
<td>4.736</td>
<td>1.310</td>
</tr>
<tr>
<td>8</td>
<td>0.005</td>
<td>2.189</td>
<td>4.755</td>
<td>1.310</td>
</tr>
<tr>
<td>9</td>
<td>0.003</td>
<td>2.189</td>
<td>4.773</td>
<td>1.310</td>
</tr>
<tr>
<td>10</td>
<td>0.002</td>
<td>2.189</td>
<td>4.791</td>
<td>1.310</td>
</tr>
</tbody>
</table>

- $P_d$ = 0.117
- $\sum \alpha_i = 0.143$
- $P_d$ = 0.378
- $\sum \alpha_i = 0.143$
- $P_d$ = 0.378
- $\sum \alpha_i = 0.951$
• System of \( n \) hospitals means optimization has \( n \) free parameters
  – Hard for to solve for large systems

• Can simplify to one-parameter problem:
  – *Theorem:* For \( F_0 = \text{N}(0, 1) \) and \( F_1 = \text{N}(\gamma, 1) \), the optimization simplifies to finding \( \mu \) to satisfy

\[
\sum_{i=1}^{n} \Phi \left( \mu - \frac{1}{\gamma} \ln(p_i) \right) = n - \kappa,
\]

and the optimal thresholds are then

\[
h_i = \mu - \frac{1}{\gamma} \ln(p_i).
\]
Consider (Hypothetical) System to Monitor 200 Largest Cities in US

- Assume probability of attack is proportional to the population in a city
• Assume
  – $2\sigma$ magnitude event
  – Constraint of 1 false signal system-wide / day

<table>
<thead>
<tr>
<th>i</th>
<th>City</th>
<th>State</th>
<th>Population</th>
<th>Pr(attack)</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>New York city</td>
<td>New York</td>
<td>8,214,426</td>
<td>0.1101</td>
<td>1.07</td>
</tr>
<tr>
<td>2</td>
<td>Los Angeles</td>
<td>California</td>
<td>3,849,378</td>
<td>0.0516</td>
<td>1.45</td>
</tr>
<tr>
<td>3</td>
<td>Chicago</td>
<td>Illinois</td>
<td>2,833,321</td>
<td>0.0380</td>
<td>1.60</td>
</tr>
<tr>
<td>4</td>
<td>Houston</td>
<td>Texas</td>
<td>2,144,491</td>
<td>0.0287</td>
<td>1.74</td>
</tr>
<tr>
<td>5</td>
<td>Phoenix</td>
<td>Arizona</td>
<td>1,512,986</td>
<td>0.0203</td>
<td>1.91</td>
</tr>
<tr>
<td>6</td>
<td>Philadelphia</td>
<td>Pennsylvania</td>
<td>1,448,394</td>
<td>0.0194</td>
<td>1.93</td>
</tr>
<tr>
<td>7</td>
<td>San Antonio</td>
<td>Texas</td>
<td>1,296,682</td>
<td>0.0174</td>
<td>1.99</td>
</tr>
<tr>
<td>8</td>
<td>San Diego</td>
<td>California</td>
<td>1,256,951</td>
<td>0.0168</td>
<td>2.01</td>
</tr>
<tr>
<td>9</td>
<td>Dallas</td>
<td>Texas</td>
<td>1,232,940</td>
<td>0.0165</td>
<td>2.01</td>
</tr>
<tr>
<td>10</td>
<td>San Jose</td>
<td>California</td>
<td>929,936</td>
<td>0.0125</td>
<td>2.16</td>
</tr>
</tbody>
</table>

• Result: $\text{Pr}(\text{signal} | \text{attack}) = 0.388$
• Naïve result: $\text{Pr}(\text{signal} | \text{attack}) = 0.283$
The graph illustrates the trade-off between the probability of detection ($P_d$) and the expected number of false signals ($\kappa$). The curve shows decreasing returns as $\kappa$ increases, indicating that as the number of false signals increases, the improvement in detection probability plateaus.

For $\kappa = 1$, the probability of detection is approximately 0.388. The line $\kappa = 1$ is marked on the x-axis.
Choosing $\gamma$ and $\kappa$

- Optimal probability of detection for various choices of $\gamma$ and $\kappa$

<table>
<thead>
<tr>
<th>$P_d$</th>
<th>$\kappa = 1$</th>
<th>$\kappa = 2$</th>
<th>$\kappa = 3$</th>
<th>$\kappa = 4$</th>
<th>$\kappa = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma = 1$</td>
<td>0.165</td>
<td>0.228</td>
<td>0.272</td>
<td>0.307</td>
<td>0.336</td>
</tr>
<tr>
<td>$\gamma = 2$</td>
<td>0.388</td>
<td>0.481</td>
<td>0.540</td>
<td>0.583</td>
<td>0.618</td>
</tr>
<tr>
<td>$\gamma = 3$</td>
<td>0.726</td>
<td>0.801</td>
<td>0.840</td>
<td>0.866</td>
<td>0.885</td>
</tr>
<tr>
<td>$\gamma = 4$</td>
<td>0.939</td>
<td>0.964</td>
<td>0.974</td>
<td>0.980</td>
<td>0.984</td>
</tr>
</tbody>
</table>

- Choice of $\kappa$ depends on available resources
- Setting $\gamma$ is subjective: what size mean increase important to detect?
Sensitivity Analyses

- **Optimal probability of detection**

  \[
  P_d \quad \begin{array}{|c|ccccc|}
  \hline
  \kappa & 1 & 2 & 3 & 4 & 5 \\
  \hline
  \gamma = 1 & 0.165 & 0.228 & 0.272 & 0.307 & 0.336 \\
  \gamma = 2 & 0.388 & 0.481 & 0.540 & 0.583 & 0.618 \\
  \gamma = 3 & 0.726 & 0.801 & 0.840 & 0.866 & 0.885 \\
  \gamma = 4 & 0.939 & 0.964 & 0.974 & 0.980 & 0.984 \\
  \hline
  \end{array}
  \]

- **Actual probability of detection**

  \[
  P_d \quad \begin{array}{|c|ccccc|}
  \hline
  \kappa & 1 & 2 & 3 & 4 & 5 \\
  \hline
  \text{Observed } \gamma = 1 & 0.137 & 0.193 & 0.235 & 0.269 & 0.298 \\
  \text{Observed } \gamma = 2 & 0.388 & 0.481 & 0.540 & 0.583 & 0.618 \\
  \text{Observed } \gamma = 3 & 0.711 & 0.790 & 0.832 & 0.859 & 0.879 \\
  \text{Observed } \gamma = 4 & 0.925 & 0.955 & 0.968 & 0.976 & 0.981 \\
  \hline
  \end{array}
  \]
Thresholds as a Function of Probability of Attack

Counties with low probability of attack $\rightarrow$ high thresholds
- Unlikely to detect attack
- Few false signals

Counties with high probability of attack $\rightarrow$ lower thresholds
- Better chance to detect attack
- Higher number of false signals
Take-Aways

• BioSense and other biosurveillance systems’ performance can be improved now at no cost
• Approach allows for customization
  – E.g., increase in probability of detection at individual location or add additional constraint to minimize false signals
• Applies to other sensor system applications:
  – Port surveillance, radiation/chem detection systems, etc.
• Details in Fricker and Banschbach (2008)
Future Research Directions

• Assess data fusion techniques for use when multiple sensors in each region
  – I.e., relax sensor (spatial) independence assumption

• Generalize from threshold detection methods to other methods that use historical information
  – I.e., relax temporal independence assumption
Biosurveillance System Optimization:

- Fricker, R.D., Jr., and D. Banschbach, Optimizing a System of Threshold Detection Sensors, in submission.

Background Information:


Detection Algorithm Development and Assessment: