Improving Biosurveillance: Optimizing Detection Thresholds

Ronald D. Fricker, Jr.
June 7, 2012
What is Biosurveillance?

• Homeland Security Presidential Directive HSPD-21 (October 18, 2007):
  – “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]

Think of Biosurveillance Like a Large System of Shewhart Control Charts

- **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]

• Each location sends data to system daily
  – Let $X_{it}$ denote residual from model predicting counts from location $i$ on day $t$
  – If no attack anywhere $X_{it} \sim F_0$ for all $i$ and $t$
  – If attack occurs on day $t$ at location $i$ then $X_{it} \sim F_1$, $t = t, t+1, ...$

• Denote probability of attack at location $i$ as $p_i$, where $\sum_i p_i = 1$

• Threshold detection: Signal on day $t^*$ if
  $$X_{it^*} \geq h_i$$
  for one or more locations
For each hospital, choice of \( h \) is compromise between probability of true and false signals.
Some Starting Assumptions

- Absent anomalies, the $X_{it}$ are independent and identically distributed (iid) according to $f_0$
- If anomaly occurs, $X_{it}$ iid according to $f_1$ for the affected data stream(s)

- That is, to start, we’re assuming the observations are independent over time and between data streams
- To achieve temporal independence, may be monitoring residuals from model that accounts for systematic effects in the data
• It’s simple to write out:

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

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

• Express it as an optimization problem:

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

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

• Absent anomalies, (standardized) data distributed according to standard normal:
  \[ F_0 = N(0,1) \]

• Anomaly manifests as a \(2\sigma\) increase in mean:
  \[ F_1 = N(2,1) \]

• Then, problem is:
  \[
  \min_h \sum_i \Phi(h_i - 2)p_i \\
  \text{s.t. } \sum_i \Phi(h_i) > n - \kappa
  \]

• Let \(n = 10\) with the following \(p_i\)s:
### 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>

<table>
<thead>
<tr>
<th>( P_d )</th>
<th>( \sum \alpha_i )</th>
<th>( \sum \alpha_i )</th>
<th>( \sum \alpha_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.117</td>
<td>0.143</td>
<td>0.143</td>
<td>0.951</td>
</tr>
</tbody>
</table>
• Monitoring $n$ data streams means optimization has $n$ free parameters (thresholds)
  – Hard for to solve for large systems

• Constraint can be expressed as an equality
  – See Fricker & Banschbach (2012) for proof: [http://faculty.nps.edu/rdfricke/frickerpa.htm](http://faculty.nps.edu/rdfricke/frickerpa.htm)

• Then can wrap the constraint into the objective function
  – Turns it into an unconstrained maximization problem
  – Unconstrained problem likely easier to solve
Specific Result Assuming Normality

• Assuming normality (and equal variances), can simplify to one-parameter problem:
  
  – Lemma: For $F_0 = \mathcal{N}(0,1)$ and $F_1 = \mathcal{N}(\gamma, 1)$, the optimization simplifies to finding $\alpha$ that satisfies

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

  and the optimal thresholds are then

  $$h_i = \alpha - \frac{1}{\gamma} \ln(p_i).$$

• See Fricker & Banschbach (2012) for derivation
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

• Result: $\Pr(\text{signal} | \text{attack}) = 0.388$
• Naïve result: $\Pr(\text{signal} | \text{attack}) = 0.283$
P_d – False Alarm Trade-Off

![Graph showing the relationship between the probability of detection (P_d) and the expected number of false signals (K'). The graph displays a curve that increases as the expected number of false signals increases, indicating a diminishing returns scenario. The point (1, 0.388) is highlighted, indicating a decreasing returns point.](image_url)
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**

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

- **Actual probability of detection**

  \[
  \begin{array}{ccccc}
  \text{P}_d & \kappa = 1 & \kappa = 2 & \kappa = 3 & \kappa = 4 & \kappa = 5 \\
  \text{Observed } \gamma = 1 & 0.137 & 0.193 & 0.235 & 0.269 & 0.298 \\
  \text{Observed } \gamma = 2 & 0.388 & 0.481 & 0.540 & 0.583 & 0.618 \\
  \text{Observed } \gamma = 3 & 0.711 & 0.790 & 0.832 & 0.859 & 0.879 \\
  \text{Observed } \gamma = 4 & 0.925 & 0.955 & 0.968 & 0.976 & 0.981 \\
  \end{array}
  \]
Optimizing a County-level System
Thresholds as a Function of Probability of Attack

Counties with low probability of attack → high thresholds
- Unlikely to detect attack
- Few false signals

Counties with high probability of attack → lower thresholds
- Better chance to detect attack
- Higher number of false signals
Relaxing the Assumptions

• Some locations may be correlated
  – E.g., hospitals in close proximity

• Example:
• Here $F_{0,i} = N(\mu_{0,i}, \Sigma_i)$ and $F_{1,i} = N(\mu_{1,i}, \Sigma_i)$ for $i=1,..,k$ groups, and we’ll assume

$$\| \mu_{0,i} - \mu_{1,i} \| = \nu$$

• For $X_i \sim F_0$

$$\left( X_{i,t} - \mu_{0,i} \right) \Sigma_i^{-1} \left( X_{i,t} - \mu_{0,i} \right) \sim \chi^2_{n_i}$$

and for $X_i \sim F_1$

$$\left( X_{i,t} - \mu_{0,i} \right) \Sigma_i^{-1} \left( X_{i,t} - \mu_{0,i} \right) \sim \chi^2_{n_i,\nu}$$

• Then, the optimal thresholds are found via

$$\max_h \sum_i \left[ 1 - \chi^2_{n_i,\nu}(h_i) \right] p_i$$

s.t. $\sum_i \left[ 1 - \chi^2_{n_i}(h_i) \right] \leq \kappa$
<table>
<thead>
<tr>
<th>Sensor $i$</th>
<th>Cluster $j$</th>
<th>$p_i$</th>
<th>Optimal (Group) Thresholds</th>
<th>“Optimal” (Individual) Thresholds</th>
<th>Adjusted “Optimal” (Individual) Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.45</td>
<td>5.81538</td>
<td>1.8946</td>
<td>1.6094</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.05</td>
<td></td>
<td>2.9931</td>
<td>2.7092</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.20</td>
<td>8.99983</td>
<td>2.9931</td>
<td>2.0150</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.05</td>
<td></td>
<td>2.9931</td>
<td>2.7092</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.05</td>
<td></td>
<td>2.9931</td>
<td>2.7092</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.10</td>
<td></td>
<td>2.6474</td>
<td>2.3627</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0.05</td>
<td>13.9231</td>
<td>2.9931</td>
<td>2.7092</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.02</td>
<td></td>
<td>3.4675</td>
<td>3.1674</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>0.02</td>
<td></td>
<td>3.4675</td>
<td>3.1674</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0.01</td>
<td></td>
<td>3.7725</td>
<td>3.5569</td>
</tr>
</tbody>
</table>

Specified $\kappa$: 0.1
Achieved $\kappa$: 0.1
Example: Probability of Detection

<table>
<thead>
<tr>
<th>Condition</th>
<th>Optimal (Group) Thresholds</th>
<th>“Optimal” (Individual) Thresholds</th>
<th>Adjusted “Optimal” (Individual) Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{ij} = 2$ for exactly one $i$ and $j$</td>
<td>0.46</td>
<td>0.39</td>
<td>0.48</td>
</tr>
<tr>
<td>$\mu_{ij} = \sqrt{2}$ for two sensors in cluster $j$</td>
<td>0.46</td>
<td>0.20</td>
<td>0.28</td>
</tr>
<tr>
<td>$\mu_{ij} = \sqrt{4/n_j}$ for all sensors in cluster $j$</td>
<td>0.46</td>
<td>0.18</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Conclusions

• Can “tune” surveillance networks using intel to improve detection performance
  – Particularly useful for surveillance networks with fixed (immovable) sensors

• Formulation explicitly accounts for allowable false signal rate
  – Failure to do so a major issue with biosurveillance

• More research required to further generalize methods
• Computer intrusion detection

• Terrorist activity detection

• Port or other perimeter security applications

✓ Most generally, monitoring set of data streams with prior information about where anomalies are likely to occur
Selected References

Biosurveillance System Optimization


SPC System Optimization


Biosurveillance Background Information

Detection Algorithm Development and Assessment


