Comparing Various Methods for Sentinel Surveillance Site Placement

Geoffrey Fairchild, Alberto Segre, Gerard Rushton, Eric Foster, Philip Polgreen

2011 ISDS
Outline

Background & Calculator

Models & Algorithms

Validation & Simulator
Background/Motivation

- Significance
  - Detect emerging diseases
  - Better understand disease mechanics
  - Assist in public health decision-making

- 19 existing influenza surveillance sites in Iowa in 2007 (network of convenience)

- Volunteer-based network of outpatient providers

- State network data is forwarded to CDC.
**Problem Statement**

*Where* should we put sites to provide the *best* surveillance?

**Formally:** Given $p$ existing surveillance sites and $n$ candidate surveillance sites, choose the $m$ best additional surveillance sites. The special case where $p=\emptyset$ is called the *de novo* formulation.
Wal-Mart

• How do we define *best* in terms of a group of sites?

• Consider Wal-Mart
  • Definition 1: near the most people
  • Definition 2: near the most disposable income
  • Definition 3: locations providing easy access to customers

• Geographic information system (GIS) research provides a suite of tools and models to solve these problems once they are well-defined.
Interactive Online Calculator

Statistics
State Population: 2930109
Existing Sites: 19 existing sites cover 1174634 people (40.09% of the population)
Selected Sites: 20 selected sites cover 1491546 people (50.90% of the population)
Total: 39 total sites cover 2488842 people (84.94% of the population)

<table>
<thead>
<tr>
<th>Existing Sites</th>
<th>Selected Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>52803</td>
<td>52402</td>
</tr>
<tr>
<td>50309</td>
<td>50613</td>
</tr>
<tr>
<td>50129</td>
<td>52245</td>
</tr>
<tr>
<td>51012</td>
<td>51104</td>
</tr>
<tr>
<td>51334</td>
<td>51502</td>
</tr>
<tr>
<td>50158</td>
<td>52001</td>
</tr>
<tr>
<td>51301</td>
<td>52349</td>
</tr>
<tr>
<td>50220</td>
<td>52601</td>
</tr>
<tr>
<td>52761</td>
<td>52501</td>
</tr>
<tr>
<td>52175</td>
<td>51041</td>
</tr>
<tr>
<td>52353</td>
<td>50548</td>
</tr>
<tr>
<td>50158</td>
<td>50401</td>
</tr>
<tr>
<td>50112</td>
<td>52742</td>
</tr>
<tr>
<td>50010</td>
<td>50674</td>
</tr>
<tr>
<td>50441</td>
<td>52057</td>
</tr>
<tr>
<td>52301</td>
<td>51455</td>
</tr>
<tr>
<td>50219</td>
<td>50563</td>
</tr>
<tr>
<td>50208</td>
<td>50423</td>
</tr>
<tr>
<td>50314</td>
<td>50022</td>
</tr>
<tr>
<td>52101</td>
<td></td>
</tr>
</tbody>
</table>
Outline

Background & Calculator

Models & Algorithms

Validation & Simulator
Maximal Coverage Model (MCM)

Choose sites in areas with the densest population.

Formally: Given a specified radius of coverage for each of the $p$ existing sites and $n$ candidate surveillance sites, choose the $m$ additional sites which maximize the population within the radius.

Note: We are using Euclidean distance and *not* travel/road distance.
MCM Variants

- **Non-Capacitated**
  - Simple: Wal-Mart can service an arbitrarily large number of people.
  - Practicality? Especially for clinics/hospitals?

- **Capacitated**
  - A more realistic model acknowledges that sites can service a limited number of people.
  - In very densely populated areas, multiple nearby Wal-Marts may be needed.
K-Median Model

Choose sites which *minimize* the *average distance* traveled by individuals to their nearest site.

**Formally**: Given each of the $p$ existing sites and $n$ candidate surveillance sites, choose the $m$ additional sites which minimize the sum of the distances from each person in a population to that person’s nearest site.
NP-Hardness

- Both the maximal coverage and K-median models are NP-hard:
  - Cannot compute the optimal solution in reasonable time.

- **Solution**: Use a *greedy approximation algorithm*:
  - Solves the problem quickly.
  - Cannot guarantee the resulting solution is optimal.
Model Summary

Maximal Coverage Model (MCM)
maximizes the population coverage within a given specified radius

K-Median Model
minimizes the average distance across the entire population
Example: Iowa Population Coverage

Population Coverage (%) vs. Number of Sites

- Existing
- MCM Capacitated
- MCM Non-Capacitated
- K-Median

computational epidemiology research

The University of Iowa
Example: Iowa Average Distance

<table>
<thead>
<tr>
<th>Existing</th>
<th>MCM Capacitated</th>
<th>MCM Non-Capacitated</th>
<th>K-Median</th>
</tr>
</thead>
</table>

- **Average Distance**
- **Number of Sites**

**Graph Details:**
- X-axis: Number of Sites
- Y-axis: Average Weighted Distance
- Lines indicate different site configurations:
  - Existing
  - MCM Capacitated
  - MCM Non-Capacitated
  - K-Median
Outline

Background & Calculator

Models & Algorithms

Validation & Simulator
Validation

**Overall Idea:** Simulate influenza spread across Iowa to compare existing sites and sites calculated *de novo*.

**In Practice:**
- Use a dataset of real influenza cases over time to replay multiple influenza seasons.
- Flip a weighted coin to determine whether each case is detected by a given site.
Iowa Medicaid Data

- Approximately 2 million cases selected from a much larger set
- Cases selected based on 30 ICD-9 codes associated with influenza-like illness (ILI)
  - Only 3 direct diagnoses of influenza
    (*influenza-specific dataset*)
- Represents 8 complete flu seasons (2000-2008)
- Medicaid billing data are *complete* with respect to its demographic (people with low incomes).
## Iowa Medicaid Data Examples

<table>
<thead>
<tr>
<th>anon_id</th>
<th>ICD-9</th>
<th>date</th>
<th>claimtype</th>
<th>patient_zip</th>
<th>age</th>
<th>gender</th>
<th>provider_zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>2421392</td>
<td>466</td>
<td>12/5/2005</td>
<td>O</td>
<td>50315</td>
<td>43</td>
<td>F</td>
<td>50314</td>
</tr>
<tr>
<td>2421392</td>
<td>465.9</td>
<td>1/23/2006</td>
<td>O</td>
<td>50315</td>
<td>43</td>
<td>F</td>
<td>50314</td>
</tr>
<tr>
<td>2421392</td>
<td>465.9</td>
<td>2/2/2006</td>
<td>O</td>
<td>50315</td>
<td>43</td>
<td>F</td>
<td>50314</td>
</tr>
<tr>
<td>2421392</td>
<td>465.9</td>
<td>3/9/2006</td>
<td>O</td>
<td>50315</td>
<td>43</td>
<td>F</td>
<td>50314</td>
</tr>
<tr>
<td>2421392</td>
<td>465.9</td>
<td>11/7/2006</td>
<td>O</td>
<td>50315</td>
<td>44</td>
<td>F</td>
<td>50314</td>
</tr>
<tr>
<td>1406011</td>
<td>780.6</td>
<td>11/30/2000</td>
<td>O</td>
<td>50316</td>
<td>37</td>
<td>F</td>
<td>50316</td>
</tr>
<tr>
<td>1979061</td>
<td>462</td>
<td>5/16/2001</td>
<td>M</td>
<td>50309</td>
<td>59</td>
<td>F</td>
<td>50315</td>
</tr>
<tr>
<td>425531</td>
<td>466</td>
<td>2/2/2004</td>
<td>M</td>
<td>50317</td>
<td>32</td>
<td>F</td>
<td>50316</td>
</tr>
<tr>
<td>425531</td>
<td>466</td>
<td>2/12/2004</td>
<td>M</td>
<td>50317</td>
<td>32</td>
<td>F</td>
<td>50313</td>
</tr>
<tr>
<td>425531</td>
<td>465.9</td>
<td>8/11/2004</td>
<td>M</td>
<td>50317</td>
<td>2</td>
<td>F</td>
<td>50313</td>
</tr>
</tbody>
</table>
Medicaid Data Visualized

- all 30 ICD-9 codes (2 million cases)
- 3 influenza-specific ICD-9 codes (30000 cases)
Medicaid Data Visualized

Number of Cases Seen for Each Distance

Distance Between Patient ZIP and Provider ZIP (miles)

log(Number of Cases)
Medicaid Data Visualized

Radius = average distance between patient ZIP and provider ZIP
Medicaid Data Visualized
Simulating Detection – Huff Score

Commonly used in GIS to model aggregate behavior.

\[ H_{ij} = \frac{A_j^\alpha D_{ij}^{-\beta}}{\sum_{j=1}^{n} A_j^\alpha D_{ij}^{-\beta}} \]

- \( H_{ij} \) = probability of case \( i \) being detected by surveillance site \( j \) conditioned on input parameters
- \( A_j \) = attractiveness of site \( j \)
- \( D_{ij} \) = distance from case \( i \) to site \( j \)
- \( \alpha \) = attractiveness enhancement parameter
- \( \beta \) = distance decay parameter
- \( n \) = total number of surveillance sites
Simulating Detection – Huff Score

- How do we measure attractiveness?
  - Wal-Mart:
    - square footage of building
    - number of items sold
    - average daily number of customers
  - We use population count inside a 20 mile radius.
    - Giving an unfair advantage to the non-capacitated MCM over capacitated.

- How do we know $\alpha$ and $\beta$?
  - We don’t.
  - Explore space $1 \leq \alpha \leq 5$ and $1 \leq \beta \leq 5$
Caution: We can only make claims within identical parameters (n, α, and β).

Examples:
  - We can compare 2 different sets of sites of size 22 with α=1 and β=1.
  - We cannot compare a set of sites of size 19 and a set of sites of size 22.
Outbreak Intensity

One way of comparing the quality of one set of sites versus another is by looking at the number (or in this case, percentage) of cases detected for each network.
Outbreak Intensity

Attractiveness Enhancement $\alpha=1$, Distance Decay $\beta=1$

Cases Detected (%) vs Number of Sites

- Existing
- MCM Capacitated
- MCM Non-Capacitated
- K-Median

complepi
computational epidemiology research

The University of Iowa
Outbreak Intensity

Attractiveness Enhancement $\alpha=3$, Distance Decay $\beta=3$

Cases Detected (%) vs Number of Sites

- Existing
- MCM Capacitated
- MCM Non-Capacitated
- K-Median
Outbreak Intensity

Attractiveness Enhancement $\alpha=5$, Distance Decay $\beta=5$

Cases Detected (%) vs. Number of Sites

- Existing
- MCM Capacitated
- MCM Non-Capacitated
- K-Median

complepi
computational epidemiology research

THE UNIVERSITY OF IOWA
Outbreak Intensity

The influenza-specific dataset generates almost identical data and allows us to draw essentially the same conclusions.
Outbreak Timing

Any evaluation of a surveillance scheme should analyze **timing**.

Ideally, we want the detected cases to reflect the true start, peak, and end of a particular disease season.

How many sites do we need in order to detect the timing of a disease season?
Outbreak Timing

Plot the ratio of cases detected vs. cases in the dataset.

If we are detecting a constant fraction of the cases, the resulting plot should be a horizontal line with slope 0.
Outbreak Timing

Histogram Ratio Considering All ICD-9 Codes (19 Sites)

Ratio

Date

Existing  MCM Non-Capacitated  MCM Capacitated  K-Median
Outbreak Timing

Histogram Ratio Considering 487.x ICD-9 Codes (19 Sites)

- Existing
- MCM Non-Capacitated
- MCM Capacitated
- K-Median
Outbreak Timing

- We need a way to compare these ratios across different sets of sites and different ICD-9 code subsets.

- We generate estimates of the parameters for a state space time series using the expectation-maximization (EM) algorithm.
  - We use the noise parameter to compare these ratios.
ICD-9 Codes

- **0**
  - **07**
    - **079**
      - **079.6**
      - **079.8**
      - **079.9**
      - **079.89**
      - **079.99**
  - **46**
    - **460**
    - **...**
  - **4**
  - **7**

- **all**
Outbreak Timing

ICD-9 Code Subsets With Prefix 0 (8 Subsets)

Number of Sites

Noise

0  (98645)
07  (98645)
079  (98645)
079.6  (10043)
079.8  (2767)
079.89  (2767)
079.9  (85835)
079.99  (85835)
Outbreak Timing

ICD-9 Code Subsets With Prefix 4 (32 Subsets)
Outbreak Timing

- There is no real pattern to the noise with respect to number of sites.
- The codes under investigation matter more than the network structure.
- **Counterintuitive result:** In Iowa, the number of sites *does not* matter for outbreak timing.
Limitations

- The focus of this paper is the topologically simple state Iowa.
- We ignore demographics when selecting sites. Medicaid data only represent a specific demographic (low income households).
- Distance measurements are Euclidean, not road/travel distance (would matter in Colorado, for example).
- All calculations concentrate a population at the centroid of a ZIP code.
Future Work

- Which ICD-9 codes are best suited for detecting intensity and timing?
- How many sites suffice to detect intensity? Timing? Both?
- What other factors affect the number of sites needed?
- Generalize results to a national level: how many, and which, sites suffice?
Thank you!

Contact Me

http://compepi.cs.uiowa.edu/index.php/Profiles/Gcfairchild

geoffrey-fairchild@uiowa.edu