Validation of masking techniques for location privacy protection of individual-level health data
Presenter Disclosures

No relationships to disclose.
Location Disclosure

Longitude: -106.686529
Latitude: 35.238214
Location Disclosure
Location Disclosure
Location Disclosure
Location Disclosure

1364 Peppoli Loop SE
Rio Rancho, NM 87124
Deaths from Katrina hit both rich, poor

An analysis of the addresses of about 545 people who died during Hurricane Katrina shows the deaths from the poor and middle class to be in proportion to the economic makeup of New Orleans.
## Geocoding Addresses

<table>
<thead>
<tr>
<th>ID</th>
<th>Address</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>123 Main St</td>
<td>12345</td>
</tr>
<tr>
<td>102</td>
<td>456 Central Ave</td>
<td>12346</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Tool of a hacker: Reverse geocoding

- Geocoding in reverse
- Relative easy, relatively new
- Key tool for “hacking” published maps

<table>
<thead>
<tr>
<th>ID</th>
<th>Address</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>123 Main St</td>
<td>12345</td>
</tr>
<tr>
<td>102</td>
<td>456 Central Ave</td>
<td>12346</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Reverse Geocoding Home locations
To Publish or Not to Publish?

FIGURE 1—Maps depicting geographic locations of travel-related disease cases by (a) hepatitis A, (b) malaria, and (c) typhoid: New York City, 2000–2006.

Figure 2 Spatial distribution of participants and CR sites in the Southern corridor of Ontario
Data Aggregation to Protect Confidentiality

locations of cases

# cases per 10,000 residents
“For publicly displayed maps ..... we randomly moved the actual location of the child within a fixed radial buffer, a technique known as jittering.”

Geographic Masking
Geographic Masking

Original + masked locations
Geographic Masking

Masked locations
Specific Aims

1. Determine the degree of privacy protection provided by geographic masking techniques

2. Determine how the degree of privacy protection of each geographic masking technique varies with masking parameters, population density and the amount of supplementary information provided

3. Determine how each geographic masking technique affects the robustness of spatial analytic methods applied to the masked data
A data release provides $k$-anonymity protection if the information for each person contained in the release cannot be distinguished from at least $k-1$ individuals whose information also appears in the release.
General Methodology

- All residential addresses
- Artificial clusters (0.5%, 2.0%, 3.0%)
- Sample cases (~1000)
sample cases (~1000) → masked cases

Spatial k-anonymity

Robustness of spatial analysis
Trade-Off in Privacy Protection

The diagram illustrates the trade-off between privacy protection and data usefulness as a function of the degree of data manipulation. As the degree of data manipulation increases from low to high, privacy protection decreases, while data usefulness increases. This relationship suggests that enhancing privacy protections may come at the cost of reduced data usefulness, and vice versa.
Software Tools

Set of custom tools developed in ArcGIS 10 using Modelbuilder and Python
Geographic Masking Techniques

- random direction, fixed radius
- random displacement within a circle
- random displacement within a donut
- Gaussian displacement
- bimodal Gaussian displacement
- location swapping within a circle
- location swapping within a donut
Scaling Masking with Population Density

scaling = \frac{1}{\sqrt{\text{PopDens}}}

Population Density Classification:
- Low
- Medium
- High
Methods – Spatial K-anonymity
Methods – Spatial K-anonymity

- Circle with radius $d$
  - Masked location
  - Original location
Methods – Spatial K-anonymity

all residential addresses

masked location

original location
Methods – Spatial K-anonymity

spatial k-anonymity = 14
Geographic Masking Techniques

- random direction, fixed radius
- Gaussian displacement
- bimodal Gaussian displacement
Results – Spatial K-anonymity

Bimodal Gaussian, uniform

K-20
K-50
Results – Spatial K-anonymity

Scaled 200m, 300m, 800m
## Results – Spatial K-anonymity

<table>
<thead>
<tr>
<th>Masking Method</th>
<th>Masking Parameter</th>
<th>% with k of 20 or less</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random direction, fixed radius</td>
<td>50 uniform</td>
<td>98.3</td>
</tr>
<tr>
<td></td>
<td>200 uniform</td>
<td>26.9</td>
</tr>
<tr>
<td></td>
<td>50/75/200</td>
<td>87.3</td>
</tr>
<tr>
<td></td>
<td>200/300/800</td>
<td>8.3</td>
</tr>
<tr>
<td>Gaussian displacement</td>
<td>50 uniform</td>
<td>95.7</td>
</tr>
<tr>
<td></td>
<td>200 uniform</td>
<td>52.7</td>
</tr>
<tr>
<td></td>
<td>50/75/200</td>
<td>86.5</td>
</tr>
<tr>
<td></td>
<td>200/300/800</td>
<td>39.2</td>
</tr>
<tr>
<td>Bimodal Gaussian displacement</td>
<td>50 uniform</td>
<td>99.7</td>
</tr>
<tr>
<td></td>
<td>200 uniform</td>
<td>42.7</td>
</tr>
<tr>
<td></td>
<td>50/75/200</td>
<td>93.4</td>
</tr>
<tr>
<td></td>
<td>200/300/800</td>
<td>17.7</td>
</tr>
</tbody>
</table>
Cluster Detection Methods

1. Clustering of aggregated data
   - Concordance with census enumeration areas
   - Spatial regression using GWR
   - Local Moran’s I and Gi*

2. Global clustering of point events
   - Quadrat analysis
   - K-nearest neighbor
   - K-functions

3. Local clustering of point events
   - Nearest neighbor hierarchical clustering
   - Spatial ellipses
   - Kernel density
   - Spatial scan statistic
Methods – Local Moran’s I

• Local Moran’s I
  – Local indicator of spatial autocorrelation
  – Identifies clusters relative to the mean

• Aggregate counts by census units

• Calculate rates

• Determine high-high clusters

• Compare results between original and masked locations
Results – Local Moran’s I

Original clusters
Methods – Kernel Density

• Dual kernel density
  – Surface from sample points
  – Surface from all residential address point
  – Divide: kernel(sample) / kernel(all)
  – Smooth to control for noise

• Boundary error analysis
  – Compare results between original and masked locations
Results – Kernel Density

- Random Direction, Fixed Radius
  - K-20 = 42.7%

- Gaussian Displacement
  - K-20 = 64.6%

- Bi-Modal Gaussian Displacement
  - K-20 = 67.2%

- No Mask

- 100, 150, 400
  - K-20 = 1.0%

- 500, 750, 2000
  - K-20 = 17.1%

  - K-20 = 1.6%
Results – Kernel Density

Kernel Density Results

K-20 = 42.7%

K-20 = 64.6%

K-20 = 67.2%

K-20 = 1.0%

K-20 = 17.1%

K-20 = 1.6%
Conclusions

• Geographic masking
  – Substantial differences in performance
  – Scaling with population density highly recommended
  – Techniques which enforce minimum displacement work best

• Spatial k-anonymity measure appears useful approach

• Clustering methods
  – Aggregated clusters quite robust (e.g. local Moran’s I)
  – Local point clustering very sensitive (e.g. kernel density)
Ongoing Research

- Comprehensive comparisons
- Tool development and sharing
- Method disclosure and reverse masking
- Multiple disclosures
Funding Support

National Science Foundation

National Institute of Environmental Health Sciences

RWJF Center for Health Policy
Publications

A list of selected publications is provided below—a complete list of all my publications can be found in my CV. Links to the online versions of the articles are provided where available.

Articles

Accuracy of TIGER Data


Dasymetric Mapping