Accuracy and Precision of the NAACCR Geocoder

Recinda L Sherman, MPH CTR
David J Lee, PhD
University of Miami,
Florida Cancer Data System
Presentation Overview

- Overview FCDS
- Overview Geocoding quality
  - Components
  - Impact on results (Florida example)
- NAACCR Geocoder Test
  - Geocoded cancer data
  - GPS well location data
  - Quality assessment
- Conclusions/Next Steps
Florida Registry—Structure

- **Florida Cancer Data System (FCDS)**
  - NPCR, Incidence based
  - Inception year 1981
  - 2nd largest cancer registry in US
    - 115,000 incident cancer cases annually
  - Contracted by University of Miami
Florida Registry—Geocoding

- Claritas/Neilson
  - Selected based on 2008 quantitative assessment
  - $$$, 185,000 cases
  - Use geocode to consolidate address
- 1981-current quarter
- No manual geocoding
  - 20% - 3% ungeocoded
  - 10% - 3% PO Box level
- Actively used
  - Cluster detection, disparities, targeting services
  - Limited quality discussion
Geocoding Quality—Components

- Match rate
  - Coverage, % with spatial location
- Precision
  - Scale
    - County center versus census block
    - NAACCR Census Tract Certainty
- Accuracy
  - Correct location
  - GIS Coordinate Quality Code
Geocoding Match

- Coverage (% matched, match type)
- Software
  - Deterministic, Probabilistic
  - Parsing algorithm, Assumptions (ties)
  - “Black box”
- Underlying street files
- Quality of address data
- Batch versus manual

Did you mean:
133 NE 2nd St, Miami,
133 SE 2nd Ave, Miami,
133 NW 2nd Ave, Miami,
133 SW 2nd St, Miami,
133 SW 2nd Ave, Miami,
Geocoding Precision

- Scale (CT Certainty, Coordinate Quality)
- GPS location
- Parcel level match
- Street level match
  - Most common
  - Interpolate along street segment
- Centroid
  - Center of polygon
    - Block, tract, zipcode, county
  - Population center, physical
Geocoding Accuracy

- Ground truth
  - GPS location
Geocoding Quality—Impact

Old Data:

Clustering of Tobacco Related Cancers
Florida 1998-2002
Female Breast Cancer

Improved Data:

Clustering of Tobacco Related Cancers
Florida 1998-2002
Invasive Female Breast Cancers

Relative Risk between 1.0 & 2.0
Relative Risk greater than 2.0

This map based on preliminary data.
To be used for intended purposes only.
Not to be reproduced or redistributed.
Created by Florida Cancer Data Systems: August 2008
Geocoding Quality—Impact

Old Data:

Improved Data:

Clustering of Tobacco Related Cancers
Florida 1998-2002
Invasive Oral Cancers

Relative Risk between 1.0 & 2.0

This map based on preliminary data.
To be used for intended purposes only.
Not to be reproduced or redistributed.
Created by Florida Cancer Data Systems.

Created by Florida Cancer Data Systems: August 2008
Test Data—Cancer

55,530 cases

- 93% Miami-Dade
  - Dx 2005-2007
  - 4% ungeocodable
  - 1% zipcode level+

- 7% Union
  - Dx 1981-current quarter
  - 17% ungeocodable
  - 75% zipcode level+
Test Data—GPS Locations

- 1,416 cases
- Drinking Water Program, well locations
  - 93% urban (Miami-Dade)
  - 7% rural (Lafayette, Union)
- GPS coordinates/associated address
  - Poorer address quality
  - Long/lat of well
    - Versus address of property with well
Match Rate (Coverage)

- **Cancer data**
  - 100% (95% vendor)
  - 100% for addresses (96% vendor)
  - 100% Union (83% vendor)

- **Well data**
  - 100%
    - all had an address
Match Type (Scale)—Cancer Data

All

- Street level 60%
  - 89% vendor
- Centroids (zip+) 40%
  - 16% vendor
  - 17% PO Box
  - 5% ungeocoded vendor

Union

- Street level 7%
  - 8% vendor
- Centroids (zip+) 92%
  - 75% zip+ vendor
  - 2% vendor
  - PO Box 62%
  - 24% vendor
  - 17% ungeocoded vendor
Match Type—Well Data

- Street level 45%
- Centroids (zip+) 55%
  - PO Box 13%
Additional metadata

- **Cancer Data**
  - 52% Exact match, 35% Soundex, 12% Relaxed match
  - **Union**
    - 98% exact match

- **Well data**
  - 47% Exact match, 35% Soundex, 17% Relaxed match
  - **Union/Lafayette**
    - 58% exact match
## Precision—CT Match

<table>
<thead>
<tr>
<th></th>
<th>Census Tract Match?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td><strong>Dade</strong></td>
<td>19.1%</td>
</tr>
<tr>
<td><strong>Union</strong></td>
<td>20.4%</td>
</tr>
</tbody>
</table>
## Precision—Cancer Data

<table>
<thead>
<tr>
<th>Polygon Type</th>
<th>County</th>
<th>Units</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>AddressRangeInterpolation</td>
<td>Miami-Dade</td>
<td>Meters</td>
<td>4203.52</td>
</tr>
<tr>
<td></td>
<td>Union</td>
<td>Meters</td>
<td>24309.94</td>
</tr>
<tr>
<td>CityCentroid</td>
<td>Miami-Dade</td>
<td>Kilometers</td>
<td>143.19</td>
</tr>
<tr>
<td></td>
<td>Union</td>
<td>Meters</td>
<td>55950396.02</td>
</tr>
<tr>
<td>CountySubdivisionCentroid</td>
<td>Miami-Dade</td>
<td>Kilometers</td>
<td>1938.33</td>
</tr>
<tr>
<td></td>
<td>Union</td>
<td>Kilometers</td>
<td>308.92</td>
</tr>
<tr>
<td>ExactParcelCentroidPoint</td>
<td>Miami-Dade</td>
<td>Meters</td>
<td>3782.98</td>
</tr>
<tr>
<td></td>
<td>Union</td>
<td>Meters</td>
<td>18725.58</td>
</tr>
<tr>
<td>StateCentroid</td>
<td>Miami-Dade</td>
<td>Kilometers</td>
<td>170310.27</td>
</tr>
<tr>
<td></td>
<td>Union</td>
<td>Kilometers</td>
<td>170310.27</td>
</tr>
<tr>
<td>ZCTACentroid</td>
<td>Miami-Dade</td>
<td>Kilometers</td>
<td>721.84</td>
</tr>
<tr>
<td></td>
<td>Union</td>
<td>Kilometers</td>
<td>21154574.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meters</td>
<td>68450888.69</td>
</tr>
</tbody>
</table>
## Precision—Wells

<table>
<thead>
<tr>
<th>Polygon Type</th>
<th>County</th>
<th>Units</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>AddressRangeInterpolation</td>
<td>DADE</td>
<td>Meters</td>
<td>6186.86</td>
</tr>
<tr>
<td></td>
<td>UNIO</td>
<td>Meters</td>
<td>14651.49</td>
</tr>
<tr>
<td>CityCentroid</td>
<td>DADE</td>
<td>Kilometers</td>
<td>143.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meters</td>
<td>31643073.35</td>
</tr>
<tr>
<td>CountySubdivisionCentroid</td>
<td>DADE</td>
<td>Kilometers</td>
<td>1938.33</td>
</tr>
<tr>
<td>ExactParcelCentroidPoint</td>
<td>DADE</td>
<td>Meters</td>
<td>12753.92</td>
</tr>
<tr>
<td>StateCentroid</td>
<td></td>
<td>Kilometers</td>
<td>170310.27</td>
</tr>
<tr>
<td>ZCTA Centroid</td>
<td>DADE</td>
<td>Kilometers</td>
<td>1739.99</td>
</tr>
<tr>
<td></td>
<td>UNIO</td>
<td>Kilometers</td>
<td>413.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meters</td>
<td>69986500.45</td>
</tr>
</tbody>
</table>
# Accuracy—Cancer Data

<table>
<thead>
<tr>
<th>Distance</th>
<th>DADE</th>
<th>UNI</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>0.4%</td>
<td>0.0%</td>
<td>0.4%</td>
</tr>
<tr>
<td>&lt;.10 mile</td>
<td>56.2%</td>
<td>4.3%</td>
<td>53.1%</td>
</tr>
<tr>
<td>&lt;.5 mile</td>
<td>28.7%</td>
<td>3.6%</td>
<td>27.2%</td>
</tr>
<tr>
<td>1 mile</td>
<td>4.1%</td>
<td></td>
<td>3.9%</td>
</tr>
<tr>
<td>5 miles</td>
<td>7.9%</td>
<td>85.8%</td>
<td>12.6%</td>
</tr>
<tr>
<td>&gt;5 miles</td>
<td>2.6%</td>
<td>6.3%</td>
<td>2.8%</td>
</tr>
</tbody>
</table>
# Accuracy—Well Data

<table>
<thead>
<tr>
<th>Distance</th>
<th>DADE</th>
<th>LAFA</th>
<th>UNIO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>0.5%</td>
<td>0.0%</td>
<td>2.3%</td>
<td>0.5%</td>
</tr>
<tr>
<td>&lt;.10 mile</td>
<td>23.4%</td>
<td>16.1%</td>
<td>18.6%</td>
<td>23.1%</td>
</tr>
<tr>
<td>&lt;.5 mile</td>
<td>29.6%</td>
<td>25.8%</td>
<td>20.9%</td>
<td>29.2%</td>
</tr>
<tr>
<td>1 mile</td>
<td>3.1%</td>
<td>9.7%</td>
<td>7.0%</td>
<td>3.4%</td>
</tr>
<tr>
<td>5 miles</td>
<td>10.4%</td>
<td>19.4%</td>
<td>39.5%</td>
<td>11.6%</td>
</tr>
<tr>
<td>&gt;5 miles</td>
<td>33.0%</td>
<td>29.0%</td>
<td>11.6%</td>
<td>32.2%</td>
</tr>
</tbody>
</table>
Conclusions

- **Quality compared to vendor**
  - Uncertain, equivalent to vendor
  - < false positives (less street level match)

- **Improvement over current vendor**
  - Additional metadata (polygon area)
  - Transparency
    - Algorithms and reference data
  - Customized for cancer registries

- **Vendor Cost**
  - $39,411.16 (2008-Jan 2012)

- **NAACCR Geocoder**
Next Steps

- **Beta version**
  - Improved underlying street data
  - Alias files
  - Additional testing
  - Volunteers?

- **Bias evaluation**
  - Age, cancer type, urban/rural
  - Comparative (vendor) and descriptive (metadata)

- **Identify useful geocoding quality metrics**
  - Standardize (intuitive)
  - Selection criteria
  - Incorporate into models