Geographic Aspects of Direct and Indirect Estimators for Small Area Estimation

Nicholas Nagle

University of Tennessee
Oak Ridge National Laboratory
Indirect Estimate
Indirect Estimate

- Borrow strength from other regions/data. Borrowing only works if you borrow from similar data.
- Smooth local estimates toward global estimates.
Direct
- Low Bias
- High Variance

Indirect
- High Bias
- Low Variance
- Lower MSE
Creating synthetic public use data with indirect methods

Goals

• Satisfy user demand for high resolution micro data

• Synthetic data should agree with other published data

• Synthetic data should have a measure of quality
Examples of High Resolution Population

Demobase
Census Bureau
Examples of High Resolution Population

Dasymetric Mapping
USGS
Examples of High Resolution Population

LandScan
Oak Ridge National Lab for DOD/DHS
What is dasymetric mapping?

Source: USGS
What is dasymetric mapping?

High Resolution Ancillary Data + Low Resolution Population Data = High Resolution Population Estimate

Source: USGS
The algorithm used for distributing people to regions is the same as that used for assigning weights to survey records.

Dasymetric Mapping or Survey Weighting, it’s all the same.

Iterative Proportional Fitting
Iterative Proportional Fitting: Visually

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sample</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Own</td>
<td>Rent</td>
</tr>
<tr>
<td>Black</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>White</td>
<td>44</td>
<td>36</td>
</tr>
<tr>
<td>Truth</td>
<td>55</td>
<td>45</td>
</tr>
</tbody>
</table>
Iterative Proportional Fitting: Mathematically

$$\max - \frac{1}{N} \sum_{it} \frac{w_{it}}{d_{it}} \log \frac{w_{it}}{d_{it}}$$

subject to

$$\sum_{i} w_{it} X_{ik} \quad = \quad X_{tk}$$

weighted sum of microdata

ancillary data region t attribute k
IPF with Measurement Error

\[ \text{max} - \frac{1}{N} \sum_{it} \frac{w_{it}}{d_{it}} \log \frac{w_{it}}{d_{it}} - \sum_{tk} \frac{e_{tk}^2}{2\sigma_{tk}^2} \]

subject to

\[ \sum_{i} w_{it} X_{ik} \]

weighted sum of microdata

\[ = X_{tk} + e_{tk} \]

ancillary data \quad ancillary error

Gaussian prior on ancillary error
Four-way Table

Homeownership x Race x Income x Census Tract
Data with Large MOEs are still usable…
In Black Neighborhoods, Black households are more likely to own than White households (after controlling for income).
Caution: pixel level uncertainties are large, but aggregations are comparable to input (by construction)
Conclusions
We need small area data.

With current funding, census tracts are too small for direct estimators.

Indirect may be more valuable than direct for tracts and other “unofficial” geographies.

When we consider indirect methods, new data types open up, including data that aren’t completely enumerated or may have their own uncertainty.
What types of benchmarks could we create from these data?
Closing thoughts

How important is direct estimation for census tracts?

The current ACS can not support tracts and smaller geographies.

We need methods that borrow strength from nearby places.

We need to rethink benchmarks. Methods exist to take advantage of incomplete ancillary data.