Investigating Solutions to Spatially Indeterminate Data: Methods of Areal Interpolation and Spatial Allocation

Matt Ruther
Urban and Public Affairs

September 30, 2015
Different Zoning Systems

ZIP tabulation areas

council districts

police districts

census tracts
Areal Interpolation Basics

- Data is often enumerated within different zoning systems (e.g., different boundaries)
- Areal interpolation is a collection of methods to convert data between zoning systems
  - Small area estimates
  - Population data or other data
- Goal of this research is to extend these methods to make them more accurate and generalizable
Prior Work

- As a method to estimate small area populations, areal interpolation is well established (Markoff & Shapiro 1973; Tobler 1979; Goodchild & Lam 1980)

- Increasingly, research is looking at ways to increase accuracy through the use of ancillary data (Eicher & Brewer 2001; Mennis & Hultgren 2006; Langford 2007; Lin, Cromley, & Zhang 2011; Qiu, Zhang, & Zhou 2012)

- The ancillary data that is used to spatially refine the estimates include land cover data (Mennis 2003; Holt, Lo, & Hodler 2004), parcel data (Tapp 2010), and street network data (Reibel and Bufalino 2005)
Example

ZIP 40210

Source Zone

9 intersecting census tracts

Target Zones
Areal Weighting

- Population within target zone is estimated as % of source zone overlap with target zone
- Based only on geography!
- Foundation for most other methods
Density Weighting

Estimated Population = 12,144

- Population “density” within target/source intersection is estimated via AW using whole target zone.
Improving Areal Interpolation

- There are other “simple” methods, but density weighting has been shown to be the most accurate.
- However, density weighting is still based on the assumption that population is evenly distributed in the target zones.
- “Intelligent” methods of areal interpolation use ancillary data to correct this issue.
Ancillary Data
Spatial Refinement Using NLCD

- Open Space Developed
- Low Density Developed
- Medium Density Developed
- High Density Developed
- Park/Greenspace
Spatial Refinement Using Street Coverage
Spatial Refinement Using Building Footprints
Spatial Refinement Using Parcels

- Industrial
- Commercial
- Residential
- Government
- Vacant/Other
Comparison of Methods (100 Deaths)

Areal Weighting

Density Weighting

Street Refined

Parcel Refined
Our Work

Modeling residential developed land in rural areas: A size-restricted approach using parcel data

Stefan Leyk*, Matt Ruther†, Barbara P. Buttenfield‡, Nicholas N. Nagle§, and Alexander K. Sturm

*Department of Geography, University of Colorado, Boulder, CO 80309, USA
†Department of Geography, University of Tennessee, Knoxville, TN 37996, USA
‡Computational Science and Engineering Division, Oak Ridge National Laboratory, USA

ABSTRACT
In most land cover data urban areas, due to different methods, land cover estimation or filtering is required. These variables can be very large, leading to the relationships between categorized on size and prediction power of the quantity prediction and development. A subspace provides strong evidence in statistical models. This is the developed land class (NUCD).

Introduction
The development of rural land has received increased research attention in recent years. As such information may contribute to a better understanding of the processes of landscape degradation, and rural changes in patterns of rural occupancy (Gee & Reddick, 2008; Wample & P indication, 2013), environmental risk assessments (Geele & Childs, 2010; Masthay & Mirov).

Comparing the effects of an NLCD-derived dasymetric refinement on estimation accuracies for multiple areal interpolation methods

Matt Ruther*, Stefan Leyk*, and Barbara P. Buttenfield

*Department of Geography, University of Colorado, Boulder, CO 80309, USA
†Department of Urban and Public Affairs, University of Louisville, Louisville, KY 40208, USA
‡Department of Geography, University of Colorado, Boulder, CO 80309, USA

(Received 14 August 2014; accepted 10 February 2015)

Compatibility among data enumerated within different time periods may be complicated by changing enumeration boundaries over time. Areal interpolation methods are commonly used to solve such issues, but are frequently based on the questionable assumption of homogenous population density within the zones. To achieve more accurate land cover or other ancillary data may be used to better characterize the underlying source zone population density surface prior to areal interpolation. Although dasymetric techniques such as these are well known, their effectiveness across different areal interpolation methods are not well established. This research compares the accuracy of a number of areal interpolation methods using temporal analysis of population data, and evaluates the ease of dasymetric mapping on interpolation accuracy. Our findings demonstrate that dasymetric refinement noticeably improves interpolation accuracy for the areal weighting, pyrometric, and target density weighting (TDW) methods of areal interpolation. A fourth method in which land cover densities are inherently incorporated, the expectation-maximization algorithm (EM), performs equally well. Our results show that the dasymetrically refined TDW method outperforms other areal interpolation methods in most instances, but suggest that the EM algorithm may be preferred as the interval between enumeration periods grows large.

Keywords: dasymetric mapping; areal interpolation; temporal analysis; spatial refinement; National Land Cover Database
Temporal Incompatibilities in Zoning Systems

1980: 89 tracts
1990: 120 tracts
2000: 354 tracts
2010: 487 tracts
Validating the Results

2000 Tract 2.00 w/2000 Blocks

2010 Tracts 2.01 and 2.02 w/2000 Blocks
<table>
<thead>
<tr>
<th>County</th>
<th>Tracts</th>
<th>Areal Weighting</th>
<th>Parcel-Refined Areal Weighting</th>
<th>Density Weighting</th>
<th>Parcel-Refined Density Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allegheny (Pittsburgh)</td>
<td>151</td>
<td>0.022</td>
<td>0.013</td>
<td>0.025</td>
<td>0.015</td>
</tr>
<tr>
<td>Clark (Las Vegas)</td>
<td>241</td>
<td>0.434</td>
<td>0.293</td>
<td>0.307</td>
<td>0.228</td>
</tr>
<tr>
<td>Hennepin (Minneapolis)</td>
<td>53</td>
<td>0.053</td>
<td>0.035</td>
<td>0.027</td>
<td>0.027</td>
</tr>
<tr>
<td>Wayne (Detroit)</td>
<td>79</td>
<td>0.064</td>
<td>0.052</td>
<td>0.037</td>
<td>0.023</td>
</tr>
</tbody>
</table>
Future Directions

- Identify ways in which parcel data and its wealth of attributes (structure size, value, built date) can be better exploited
- Incorporate alternative ancillary data types, such as census tract/block attributes, into the interpolation
- Evaluate area interpolation methods in the context of public health data
- *Validate* the interpolated public health data

- Takeaway....
Spatial Allocation of Microdata

Microdata (NCHS/PUMS)  
*Individuals*
- Coarse geographic scale
- Extensive demographic detail

Summary Data  
*Tracts (or sub-county areas)*
- Fine geographic scale
- Limited demographic detail
Probabilistically impute new weights for each PUMS record for each of the tracts within the PUMA/county, based on the known populations of the tracts and some attributes (constraining variables) of the individual.

Does not “place” individuals!
Maximum Entropy Estimation

\[
\max \sum_i \sum_j (w_{ij}) \log \left( \frac{w_{ij}}{d_{ij}} \right) \quad \text{subject to} \quad \sum_i w_{ij}x_{ik} = X_{jk}
\]

\[i = \text{individual}\]
\[j = \text{tract}\]
\[k = \text{attribute}\]
\[d = \text{initial sampling weight}\]
\[w = \text{imputed sampling weight}\]
\[x = \text{individual demographic characteristics}\]
\[X = \text{tract aggregate demographic characteristics}\]
Prior Research

- Reweighting: Statistically adjusting the sampling weights for each HH in a survey to fit a known population distribution (Johnston & Pattie 1993; Mrozinski & Cromley 1999; Simpson & Tranmer 2005; Ballas et al. 2005)

- Complementary topic in geography is dasymetric mapping (Semenov-Tian-Shansky 1928; Wright 1936; Eicher & Brewer 2001; Mennis 2006; Riebel & Agrawal 2007)

- Much research on Census microdata reweighting has focused on UK and Australia – generally, lack 100% validation (Johnston & Pattie 1993; Williamson, Birkin, & Rees 1998; Melhuish, Blake, & Day 2002; Ballas et al. 2005; Smith, Clarke, & Harland 2009)
Goals of the Research

- Small area estimates useful in the analysis of sociodemographic processes at the local level (e.g., public health, transportation, emergency planning)
- These estimates may be used to assess the needs for schools, parks, public transportation, and health-prevention programs, and to evaluate the impact of public policies
- While some of these estimates can be made with a survey instrument, most others would need to rely on population estimation methods
- Is there ANY utility to this method in the context of health data?
Study Area and Data

- Mortality data from NCHS for 2000-2003
- Tract-level data from Census for 2000
## County Population (2000)

<table>
<thead>
<tr>
<th>County</th>
<th>Total</th>
<th>% 75+</th>
<th>% Male</th>
<th>% Black</th>
<th>% Hisp</th>
<th>Tracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>333,219</td>
<td>3</td>
<td>50</td>
<td>11</td>
<td>27</td>
<td>85</td>
</tr>
<tr>
<td>Arapahoe</td>
<td>454,271</td>
<td>4</td>
<td>51</td>
<td>15</td>
<td>11</td>
<td>121</td>
</tr>
<tr>
<td>Boulder</td>
<td>273,758</td>
<td>4</td>
<td>51</td>
<td>7</td>
<td>10</td>
<td>68</td>
</tr>
<tr>
<td>Denver</td>
<td>516,902</td>
<td>6</td>
<td>50</td>
<td>19</td>
<td>30</td>
<td>136</td>
</tr>
<tr>
<td>Jefferson</td>
<td>493,797</td>
<td>5</td>
<td>50</td>
<td>7</td>
<td>9</td>
<td>133</td>
</tr>
<tr>
<td>Weld</td>
<td>166,893</td>
<td>4</td>
<td>50</td>
<td>5</td>
<td>26</td>
<td>37</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>County</th>
<th>Total</th>
<th>% 75+</th>
<th>% Male</th>
<th>% Black</th>
<th>% Hisp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>6,447</td>
<td>46</td>
<td>50</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Arapahoe</td>
<td>8,378</td>
<td>56</td>
<td>48</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Boulder</td>
<td>4,257</td>
<td>60</td>
<td>45</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Denver</td>
<td>13,334</td>
<td>55</td>
<td>50</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Jefferson</td>
<td>9,710</td>
<td>58</td>
<td>48</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Weld</td>
<td>3,472</td>
<td>55</td>
<td>50</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Create 56 groupings determined by gender (male/female), race (black/non-black), ethnicity (Hispanic/non-Hispanic), and age (<35, 35-44, 45-54, 55-64, 65-74, 75-84, 85+)

Generate synthetic living population based on Census count of population and deaths during 2000-2003
<table>
<thead>
<tr>
<th>M</th>
<th>B</th>
<th>H</th>
<th>A</th>
<th>D</th>
<th>Tract 1</th>
<th>Tract 2</th>
<th>Tract 3</th>
<th>...</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>73</td>
<td>1</td>
<td>0.0055</td>
<td>0.0062</td>
<td>0.0078</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>59</td>
<td>0</td>
<td>0.0055</td>
<td>0.0062</td>
<td>0.0078</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>72</td>
<td>0</td>
<td>0.0055</td>
<td>0.0062</td>
<td>0.0078</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>81</td>
<td>1</td>
<td>0.0055</td>
<td>0.0062</td>
<td>0.0078</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>35</td>
<td>0</td>
<td>0.0055</td>
<td>0.0062</td>
<td>0.0078</td>
<td>...</td>
</tr>
</tbody>
</table>

Total: 2,850 3,228 4,047 ... 516,902

<table>
<thead>
<tr>
<th>M</th>
<th>B</th>
<th>H</th>
<th>A</th>
<th>D</th>
<th>Tract 1</th>
<th>Tract 2</th>
<th>Tract 3</th>
<th>...</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>73</td>
<td>1</td>
<td>0.0039</td>
<td>0.0060</td>
<td>0.0047</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>59</td>
<td>0</td>
<td>0.0053</td>
<td>0.0070</td>
<td>0.0052</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>72</td>
<td>0</td>
<td>0.0030</td>
<td>0.0068</td>
<td>0.0198</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>81</td>
<td>1</td>
<td>0.0021</td>
<td>0.0027</td>
<td>0.0058</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>35</td>
<td>0</td>
<td>0.0036</td>
<td>0.0054</td>
<td>0.0113</td>
<td>...</td>
</tr>
</tbody>
</table>

Total: 2,850 3,228 4,047 ... 516,902
Validation

- Tract-level mortality counts by age, sex, race, and ethnicity from Colorado Department of Public Health
- Compare actual counts to allocated counts on a number of tract-level (CV) and aggregate-level (RMSE) metrics
- Assess spatial patterns in the accuracy of the allocation, to improve model
## Validation Results

### Denver County (135 tracts)

<table>
<thead>
<tr>
<th>Measure</th>
<th>All</th>
<th>Cancer</th>
<th>Heart</th>
<th>Stroke</th>
<th>Diabetes</th>
<th>Flu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths</td>
<td>13,334</td>
<td>2,857</td>
<td>3,020</td>
<td>762</td>
<td>319</td>
<td>285</td>
</tr>
<tr>
<td>Spearman</td>
<td>0.86</td>
<td>0.81</td>
<td>0.82</td>
<td>0.67</td>
<td>0.50</td>
<td>0.56</td>
</tr>
<tr>
<td>MRAD</td>
<td>0.23</td>
<td>0.28</td>
<td>0.29</td>
<td>0.51</td>
<td>0.73</td>
<td>0.66</td>
</tr>
</tbody>
</table>

### Total Metropolitan Area (576 tracts)

<table>
<thead>
<tr>
<th>Measure</th>
<th>All</th>
<th>Cancer</th>
<th>Heart</th>
<th>Stroke</th>
<th>Diabetes</th>
<th>Flu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths</td>
<td>45,598</td>
<td>10,192</td>
<td>10,294</td>
<td>2,811</td>
<td>1,042</td>
<td>1,015</td>
</tr>
<tr>
<td>Spearman</td>
<td>0.90</td>
<td>0.84</td>
<td>0.86</td>
<td>0.74</td>
<td>0.49</td>
<td>0.61</td>
</tr>
<tr>
<td>MRAD</td>
<td>0.24</td>
<td>0.27</td>
<td>0.33</td>
<td>0.51</td>
<td>0.85</td>
<td>0.72</td>
</tr>
</tbody>
</table>
Validation Results (Cause-Specific)

Cancer Deaths

Flu Deaths

Relative Absolute Deviation
- 0.00 - 0.25
- 0.26 - 0.50
- 0.51 - 0.75
- 0.76 - 1.00
- 1.01 +

No Deaths
Future Directions

- Does it work?!
- How to incorporate additional constraints?
- Improve model by combining similar tracts?
- Evaluate the use of morbidity data (additional problems....)
Investigating Solutions to Spatially Indeterminate Data: Methods of Areal Interpolation and Spatial Allocation

Matt Ruther
Urban and Public Affairs
matthew.ruther@louisville.edu

September 30, 2015