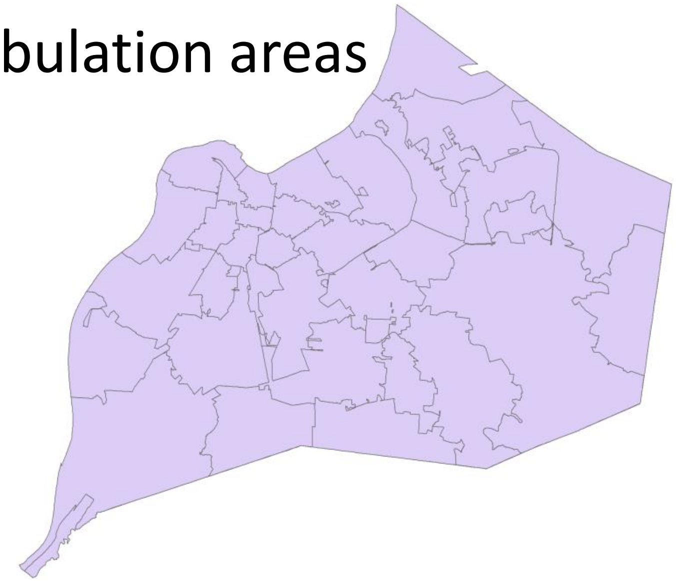
Investigating Solutions to Spatially Indeterminate Data: Methods of Areal Interpolation and Spatial Allocation

Matt Ruther
Urban and Public Affairs

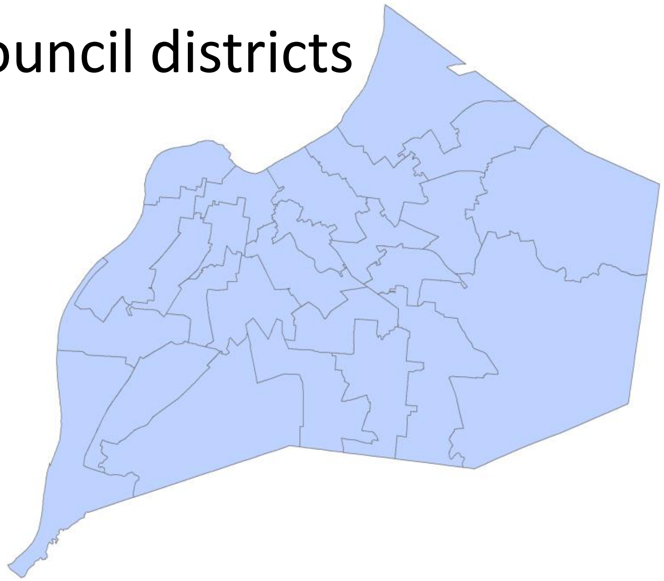
September 30, 2015

Different Zoning Systems

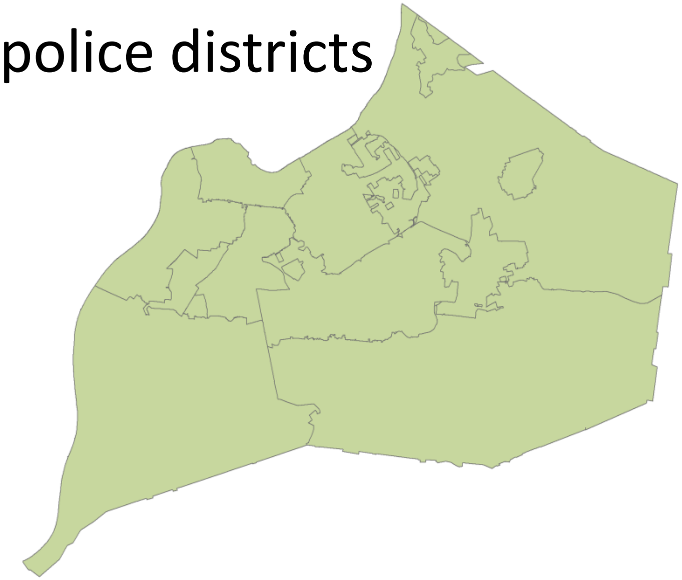
ZIP tabulation areas



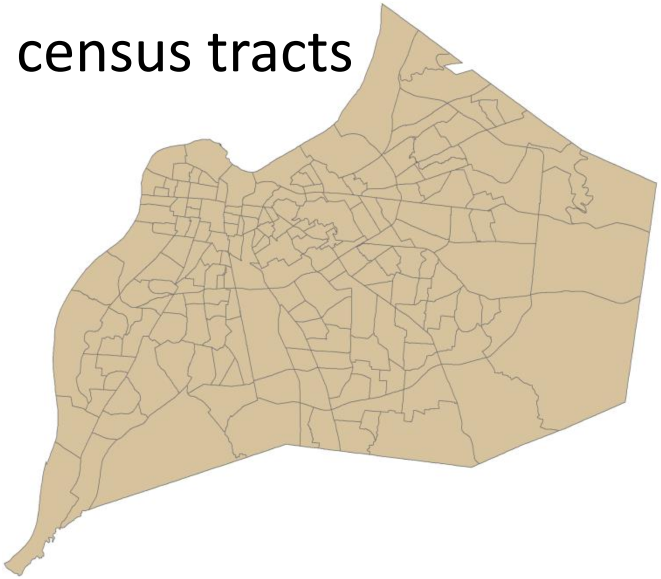
council districts



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

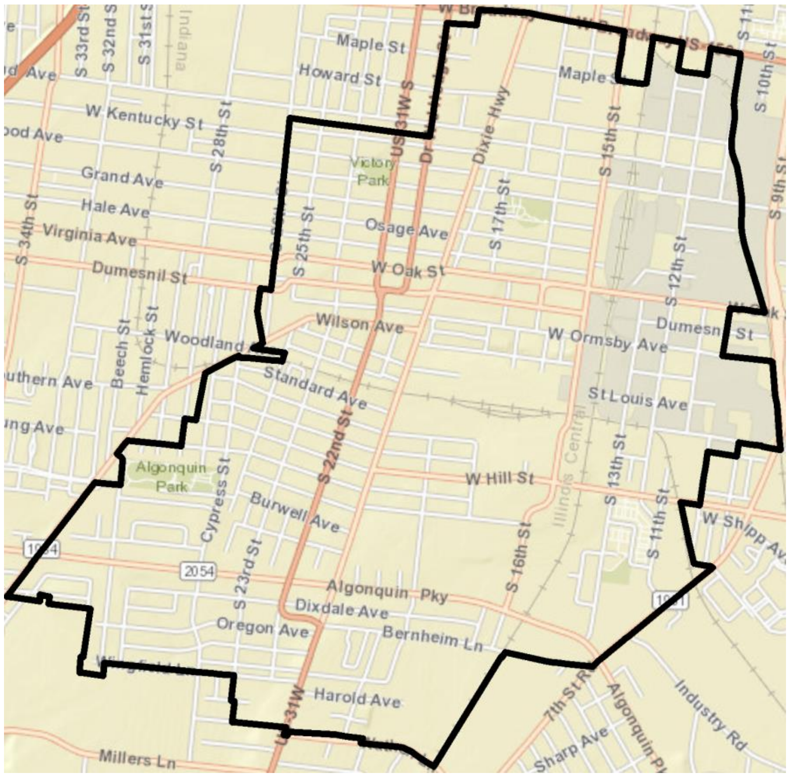
- As a method to estimate small area populations, areal interpolation is well established (Markoff & Shapiro 1973; Tobler 1979; Goodchild & Lam 1980)

- 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)

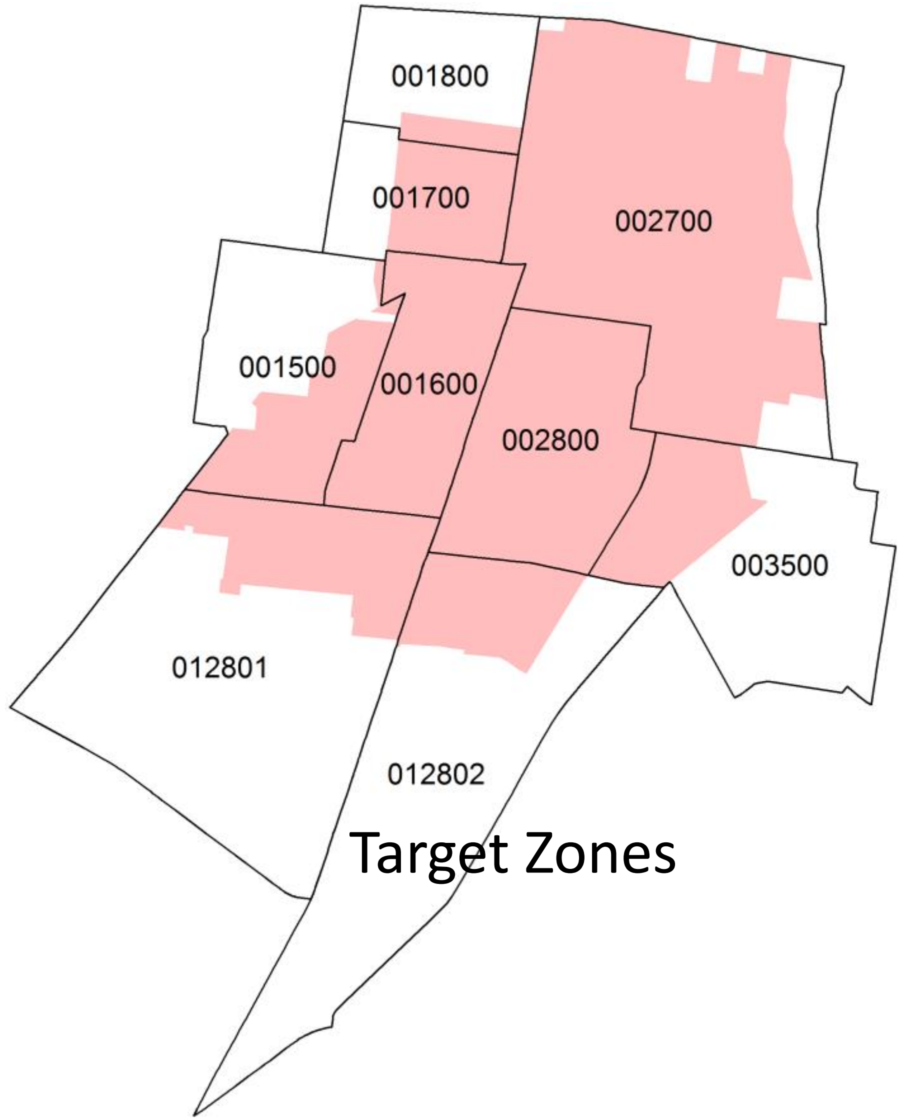
Example

ZIP 40210



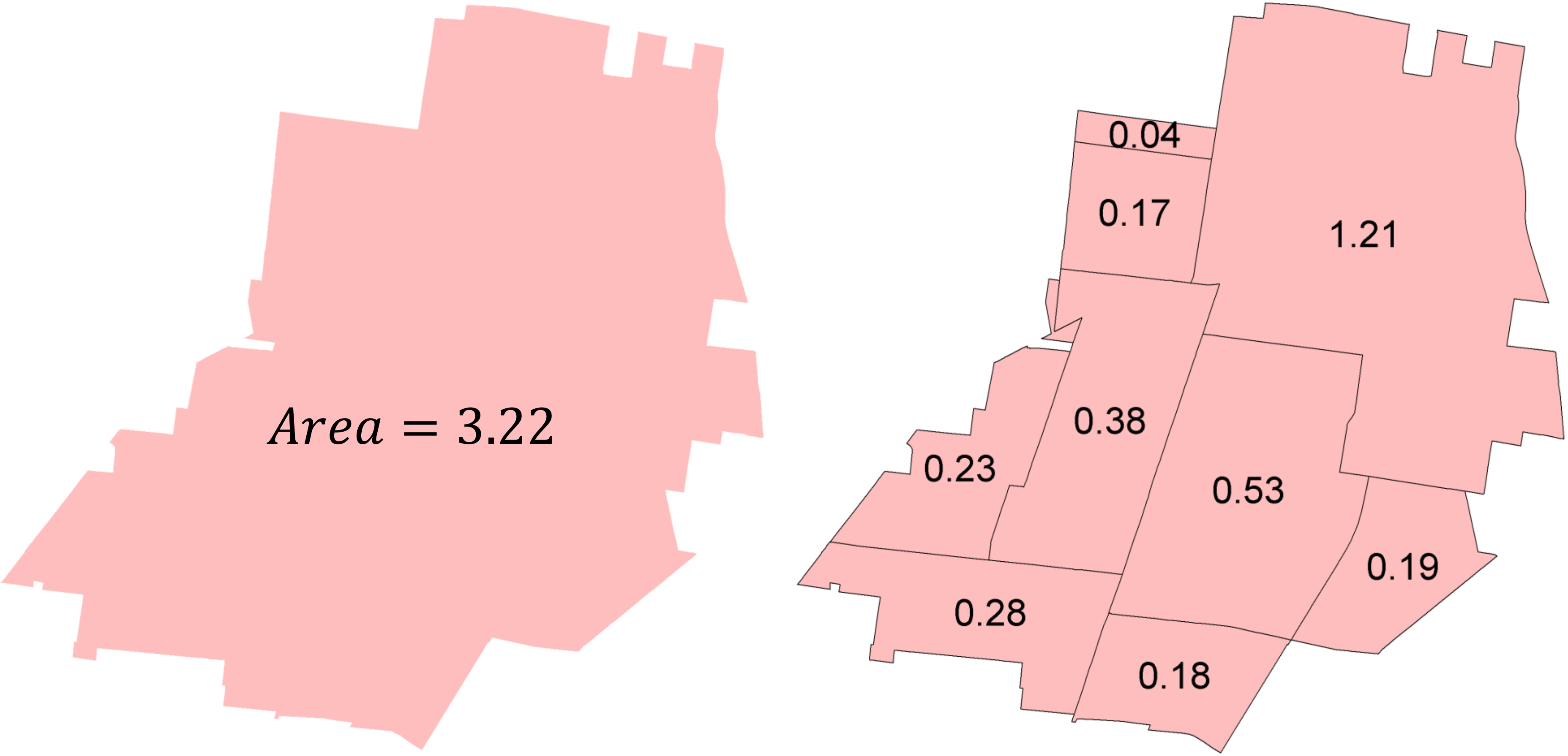
Source Zone

9 intersecting census tracts



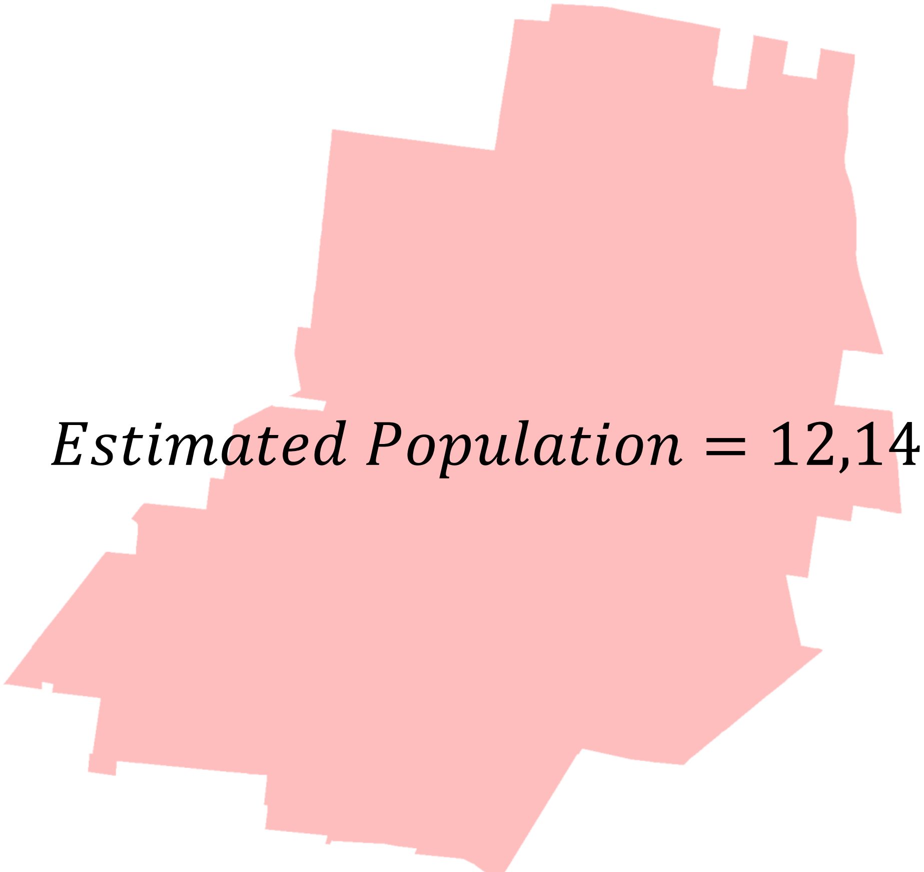
Target Zones

Areal Weighting

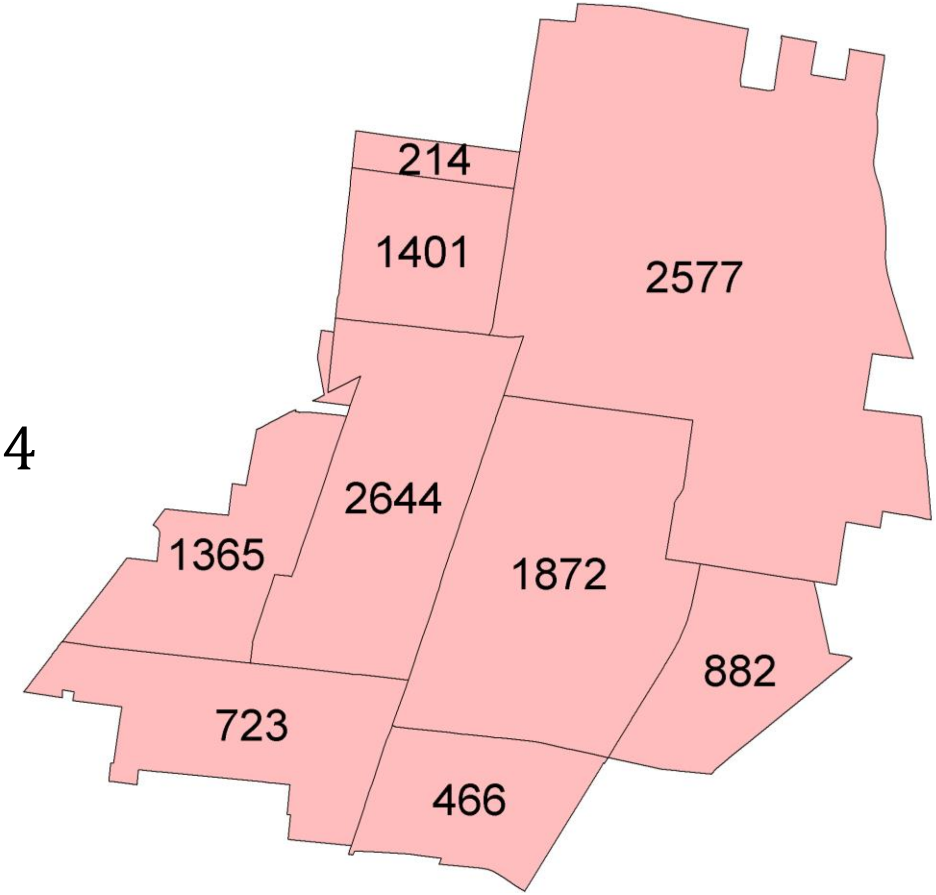


- Population within target zone is estimated as % of source zone overlap with target zone
- Based only on geography!
- Foundation for most other methods

Density Weighting



$$Estimated\ Population = 12{,}144$$

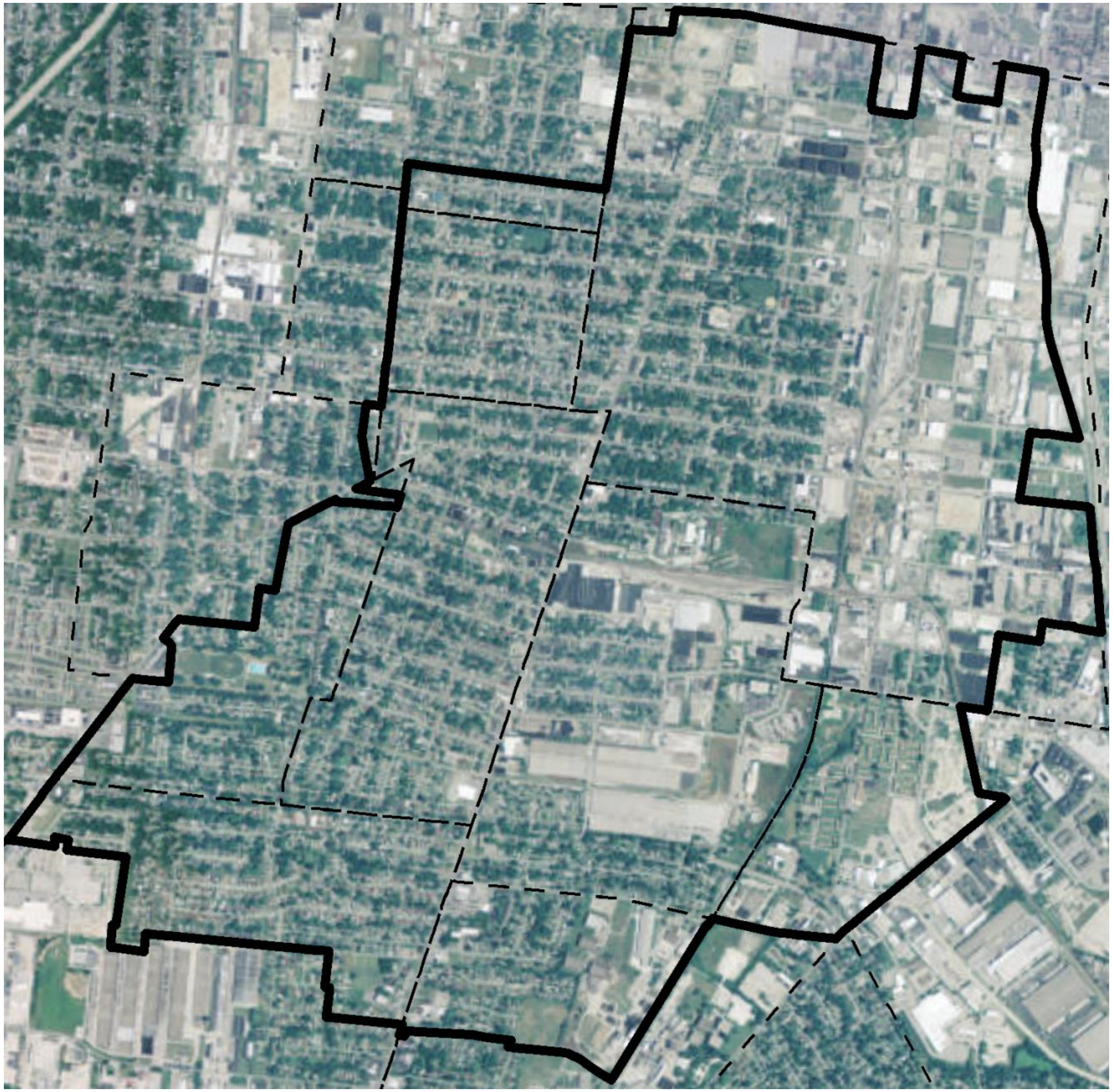


- Population “density” within target/source intersection is estimated via AW using whole target zone

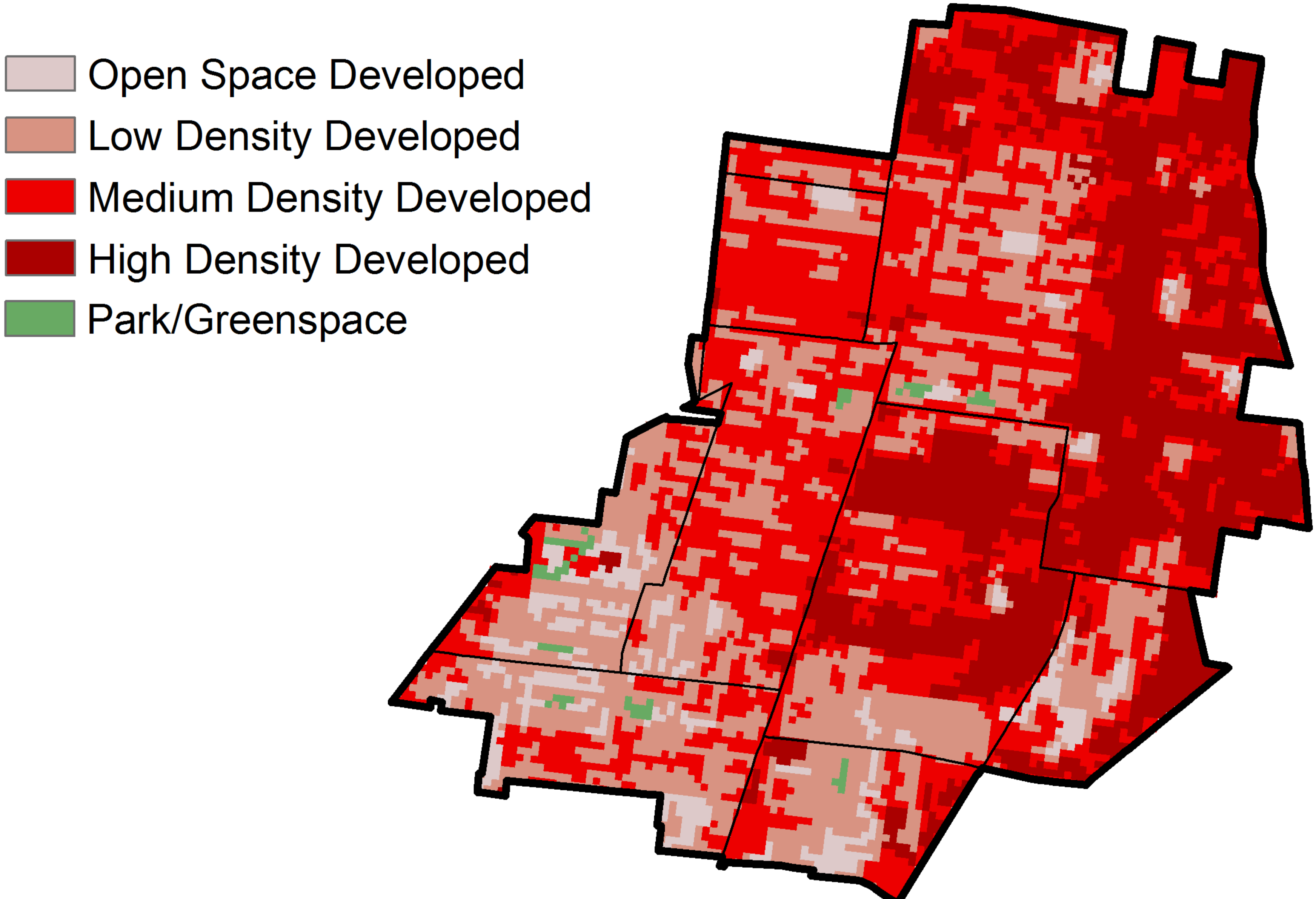
Improving Areal Interpolation

- 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

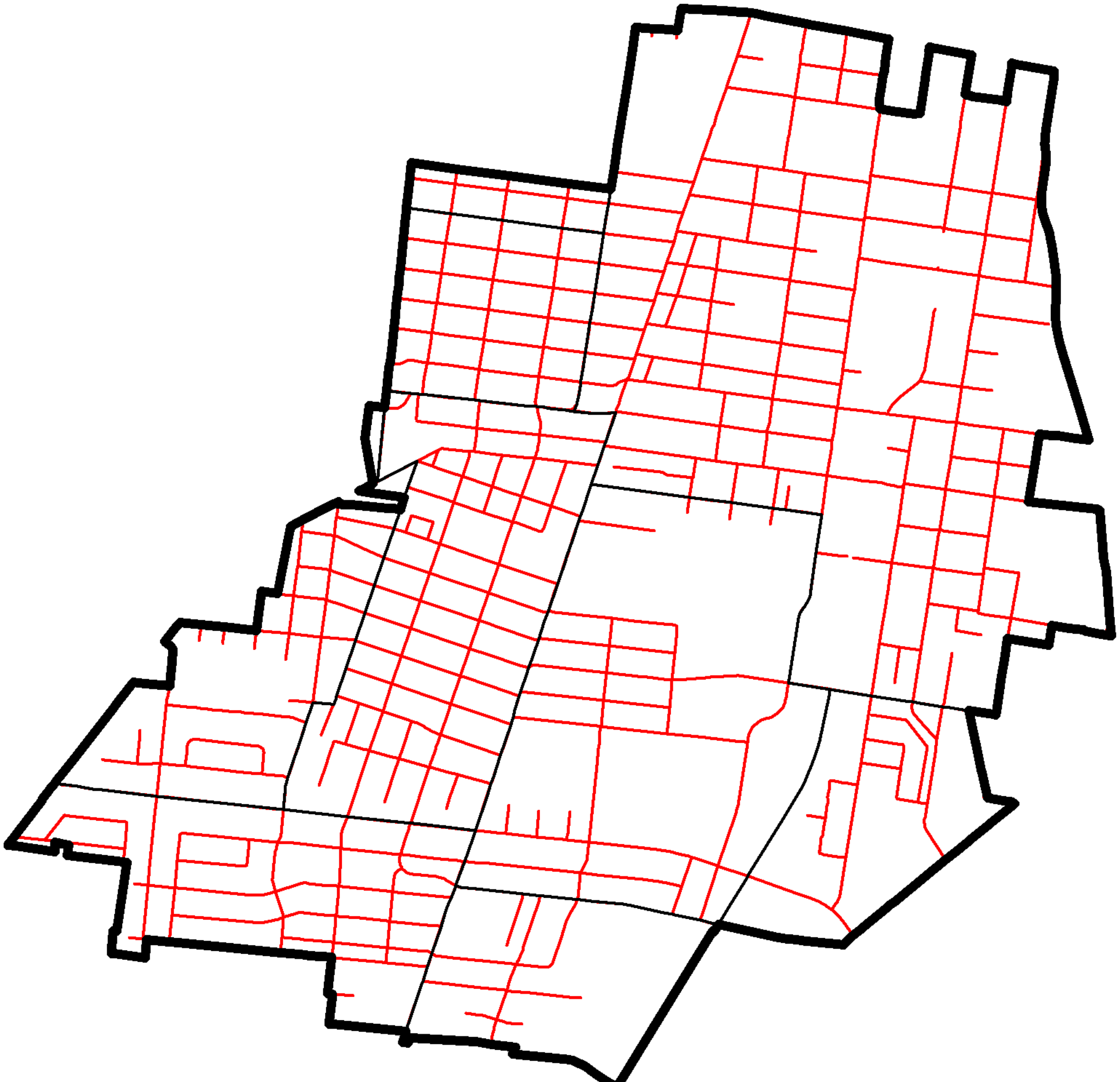
Ancillary Data



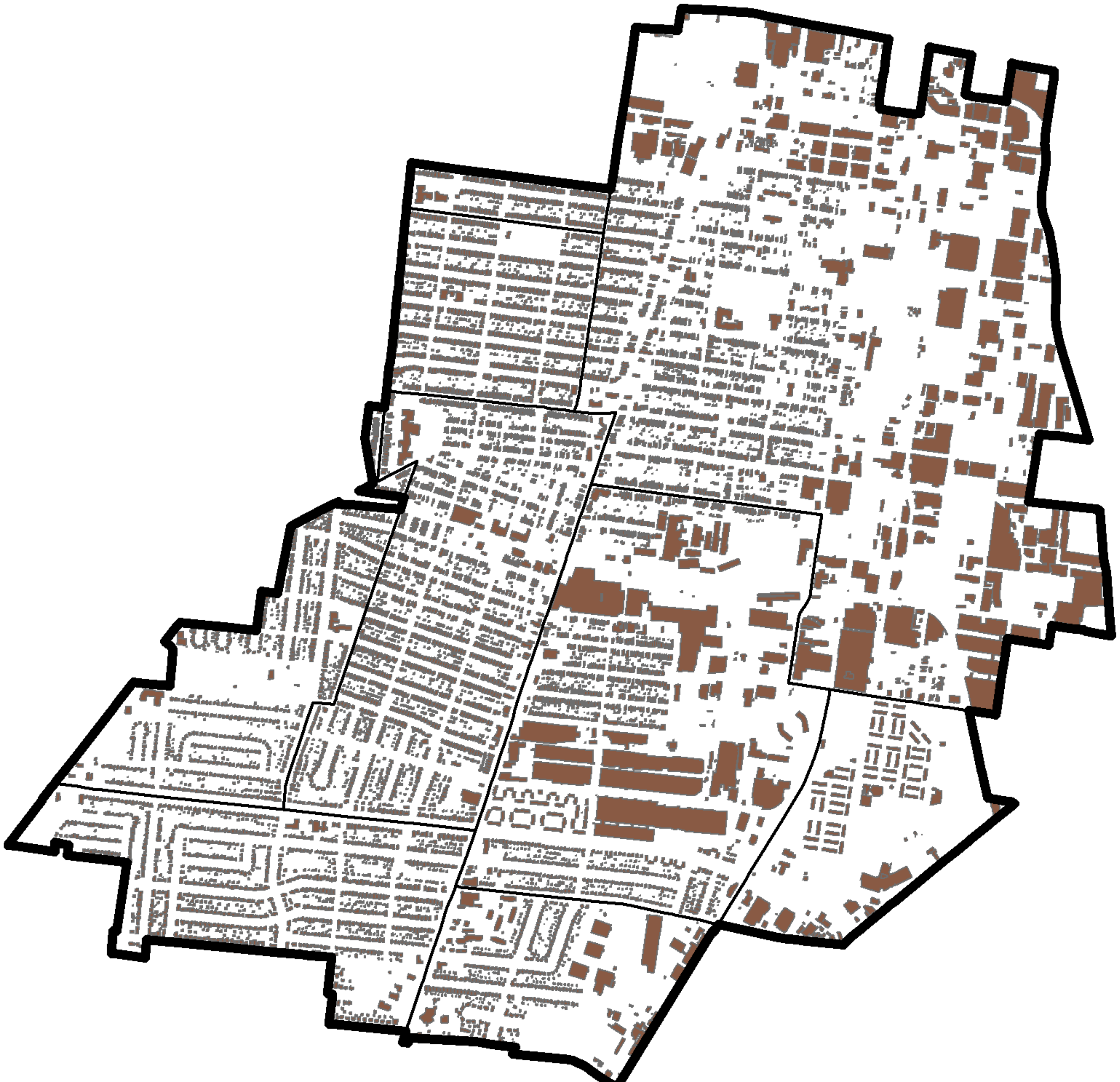
Spatial Refinement Using NLCD



Spatial Refinement Using Street Coverage

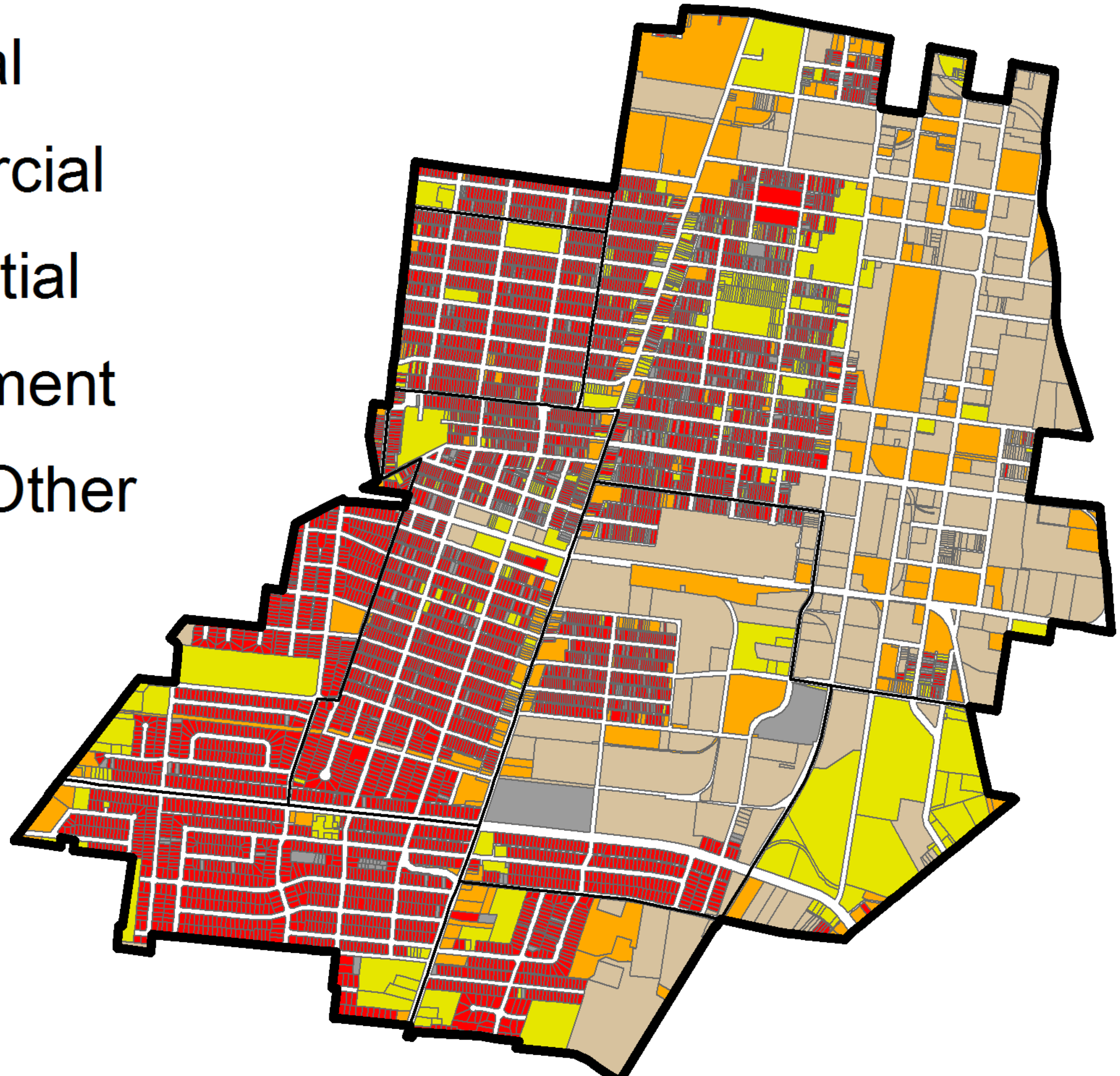


Spatial Refinement Using Building Footprints

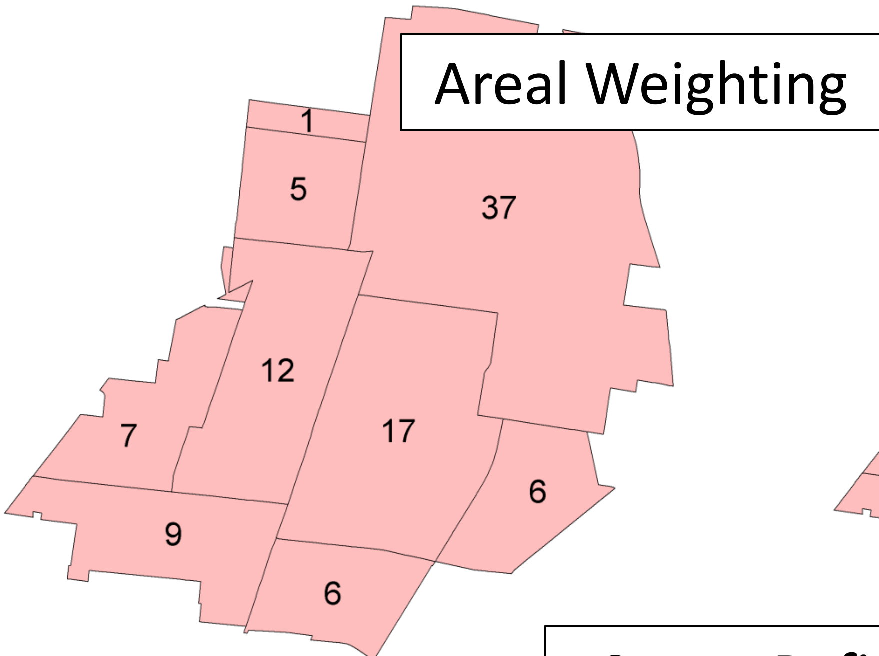


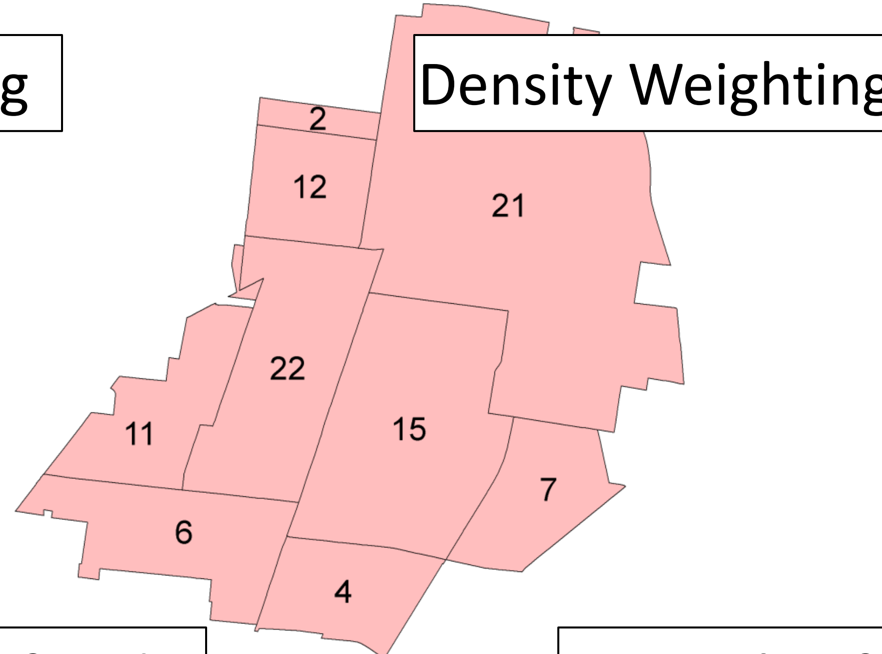
Spatial Refinement Using Parcels

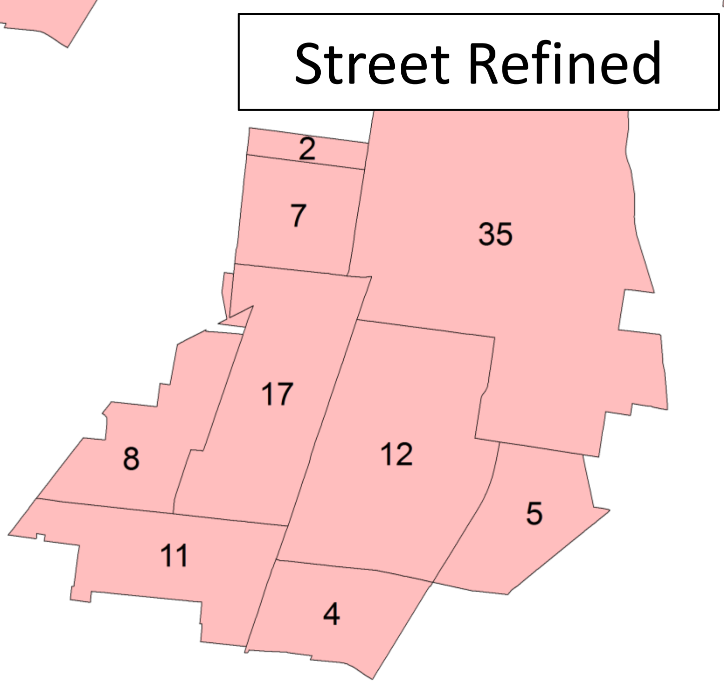
Industrial
Commercial
Residential
Government
Vacant/Other

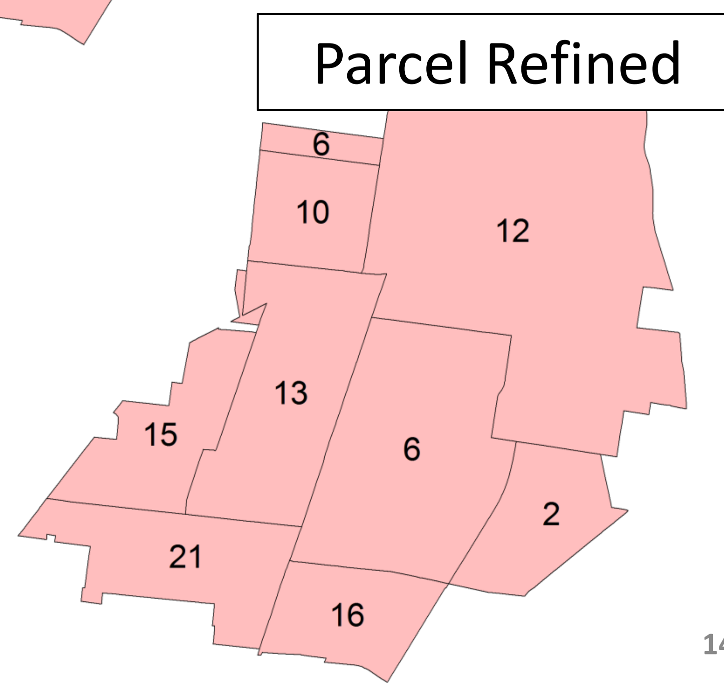


Comparison of Methods (100 Deaths)









Our Work

Applied Geography 47 (2014) 33–45

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

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

Stefan Leyk [a,*], Matt Ruther [a], Barbara P. Buttenfield [a], Nicholas N. Nagle [b,c], Alexander K. Stum [a]

[a] Department of Geography, University of Colorado, Boulder, CO 80309, USA
[b] Department of Geography, University of Tennessee, Knoxville, TN 37996, USA
[c] Computational Sciences and Engineering Division, Oak Ridge National Laboratory, USA

Keywords:
Small area estimation
Rural areas
Developed land
Land cover
Dasymetric mapping

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 fragmentation and urbanization, as well as changes in patterns of rural occupancy (Irwin & Bockstael, 2007; Irwin, Cho, & Bockstael, 2007). Advanced knowledge of such processes has direct implications for research in problem domains such as demographic small area estimation (Mennis, 2009; Zandbergen & Ignizio, 2011), public service accessibility (Langford, Higgs, Radcliffe, & White, 2008), environmental risk assessments (Giordano & Cheever, 2010; Maantay & Maroko,

* Corresponding author. Department of Geography, University of Colorado, UCB 260, Boulder, CO, USA.
E-mail addresses: stefan.leyk@colorado.edu (S. Leyk), matthew.ruther@colorado.edu (M. Ruther), babs@colorado.edu (B.P. Buttenfield), nnagle@utk.edu (N.N. Nagle), alexander.stum@colorado.edu (A.K. Stum).

Cartography and Geographic Information Science, 2015
http://dx.doi.org/10.1080/15230406.2015.1065206

Taylor & Francis

Exploring the impact of dasymetric refinement on spatiotemporal small area estimates

Barbara P. Buttenfield [a*], Matt Ruther [b] and Stefan Leyk [a]

[a] Department of Geography, University of Colorado, Boulder, CO, USA; [b] Department of Urban and Public Affairs, University of Louisville, Louisville, KY, USA

(Received 29 August 2014; accepted 3 April 2015)

GIScience & Remote Sensing, 2015
Vol. 52, No. 2, 158–178, http://dx.doi.org/10.1080/15481603.2015.1018856

Taylor & Francis

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

Matt Ruther [a*], Stefan Leyk [b] and Barbara P. Buttenfield [b]

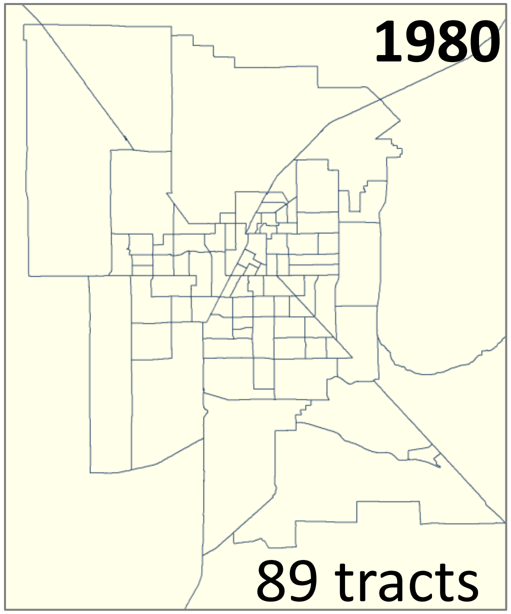
[a] Department of Urban and Public Affairs, University of Louisville, Louisville, KY 40208, USA;
[b] Department of Geography, University of Colorado Boulder, Boulder, CO 80309, USA

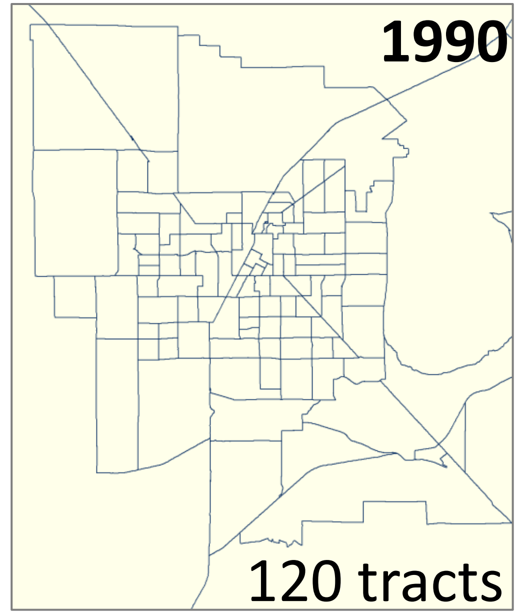
(Received 14 August 2014; accepted 10 February 2015)

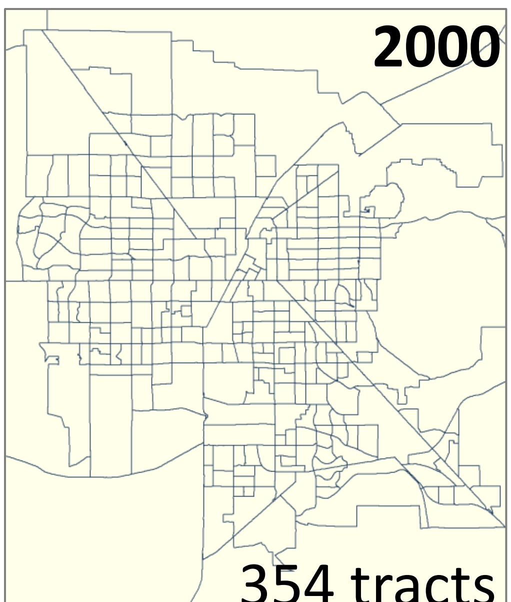
Comparability among population data enumerated within different time periods may be complicated by changing enumeration boundaries over time. Areal interpolation methods are commonly used to solve such zoning incompatibilities, but are frequently based on the questionable assumption of homogeneous population density within the different zones. To achieve more accurate estimates, 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 documented, their effectiveness across different areal interpolation methods are not well established. This research compares the accuracy of a number of areal interpolation methods used to support temporal analysis of population data, and evaluates the effect of dasymetric mapping on interpolation accuracy. Our findings demonstrate that dasymetric refinement noticeably improves interpolation accuracy for the areal weighting, pycnophylactic, 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

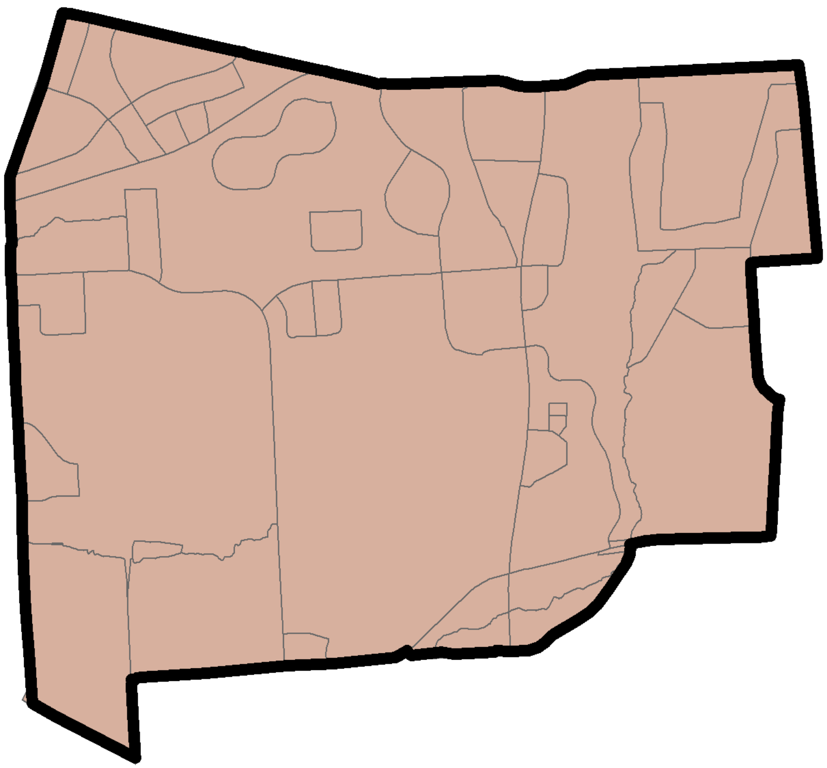




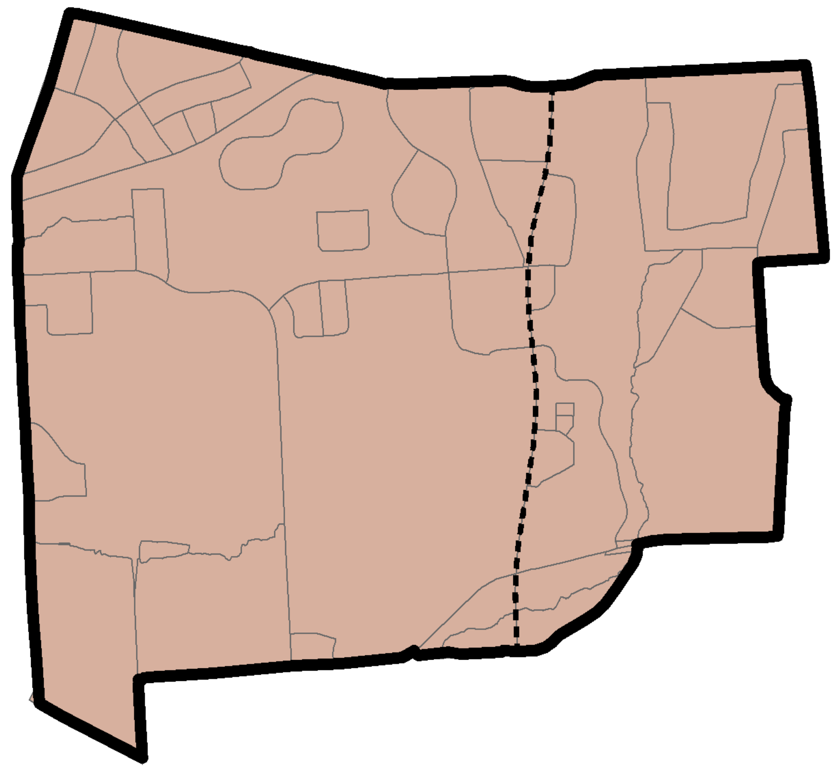


Validating the Results

2000 Tract 2.00
w/2000 Blocks



2010 Tracts 2.01 and 2.02
w/2000 Blocks



Median Standardized Absolute Error by County and Interpolation Method

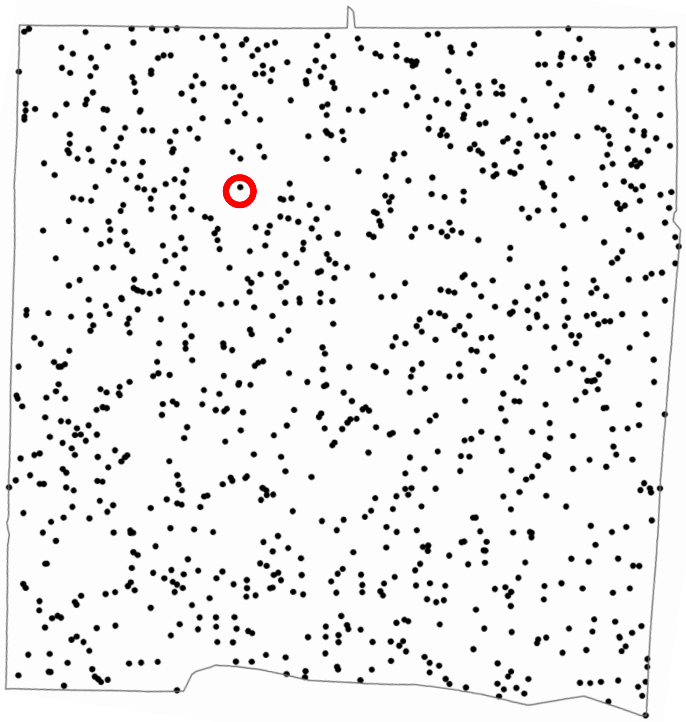
County	Tracts	Areal Weighting	Parcel-Refined Areal Weighting	Density Weighting	Parcel-Refined Density Weighting
Allegheny (Pittsburgh)	151	0.022	0.013	0.025	0.015
Clark (Las Vegas)	241	0.434	0.293	0.307	0.228
Hennepin (Minneapolis)	53	0.053	0.035	0.027	0.027
Wayne (Detroit)	79	0.064	0.052	0.037	0.023

Future Directions

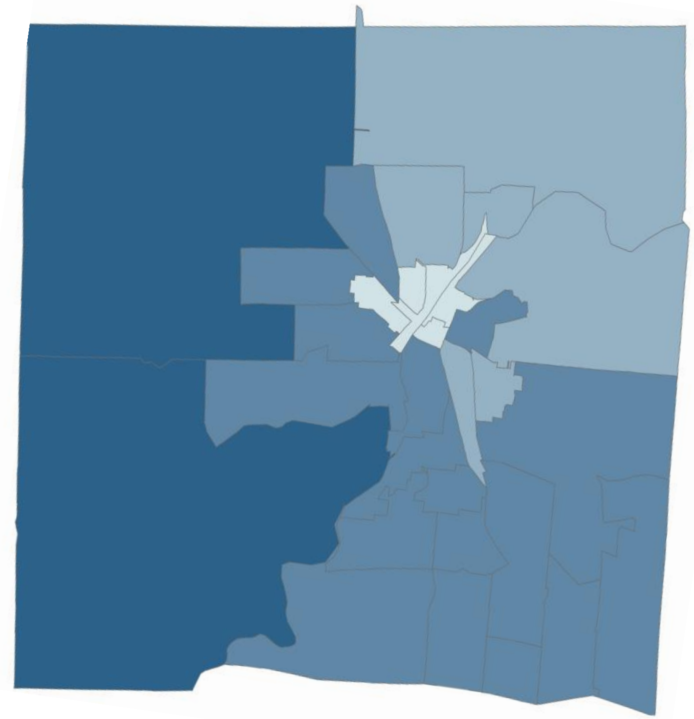
- Identify ways in which parcel data and its wealth of attributes (structure size, value, built date) can be better exploited
- Incorporate alternative ancillary data types, such as census tract/block attributes, into the interpolation
- Evaluate area interpolation methods in the context of public health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

- Takeaway….

Spatial Allocation of Microdata

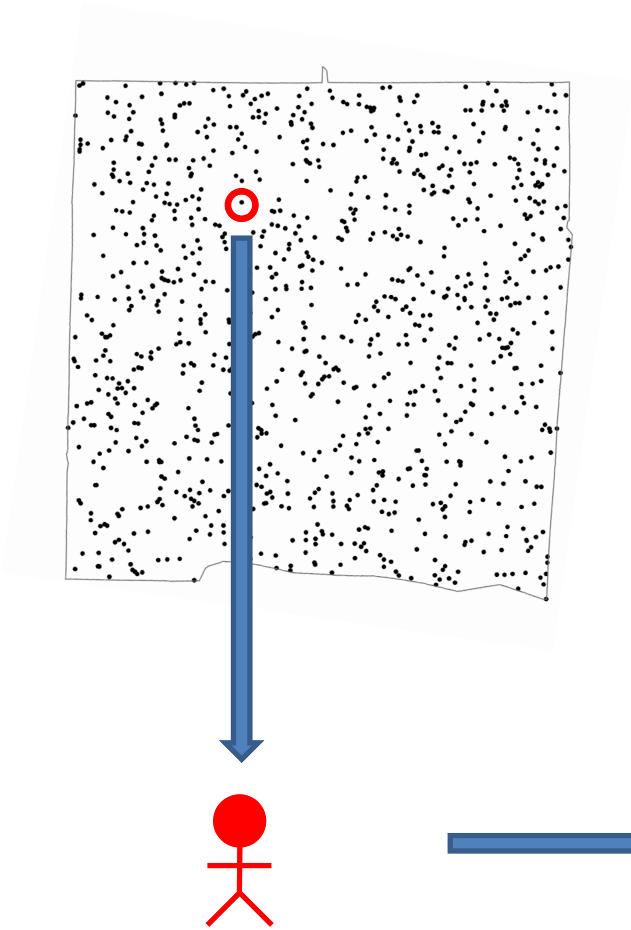


Microdata (NCHS/PUMS)
Individuals
Coarse geographic scale
Extensive demographic detail



Summary Data
Tracts (or sub-county areas)
Fine geographic scale
Limited demographic detail

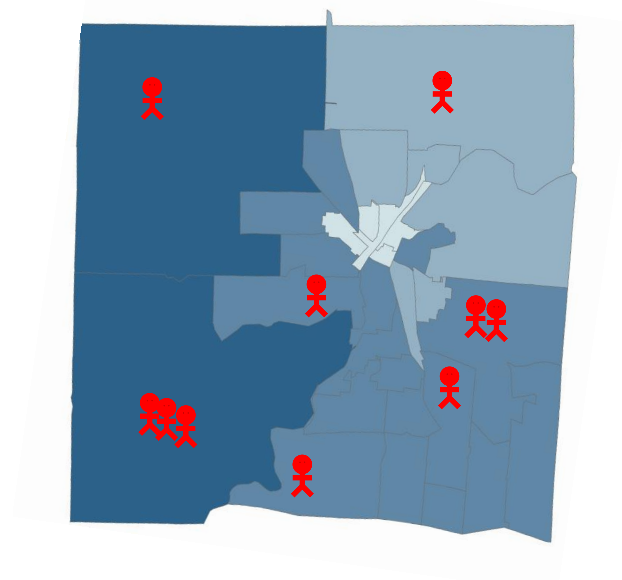
In Pictures



1 NCHS/PUMS Record
(Weight = 10)

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

$$\boxed{max \sum_i \sum_j (w_{ij}) \, log\left(\frac{w_{ij}}{d_{ij}}\right) \text{ subject to } \sum_i w_{ij} x_{ik} = X_{jk}}$$

$i =$ $individual$

$j =$ $tract$

$k =$ $attribute$

$d =$ $initial\ sampling\ weight$

$w =$ $imputed\ sampling\ weight$

$x =$ $individual\ demographic\ characteristics$

$X =$ $tract\ aggregate\ demographic\ characteristics$

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)*

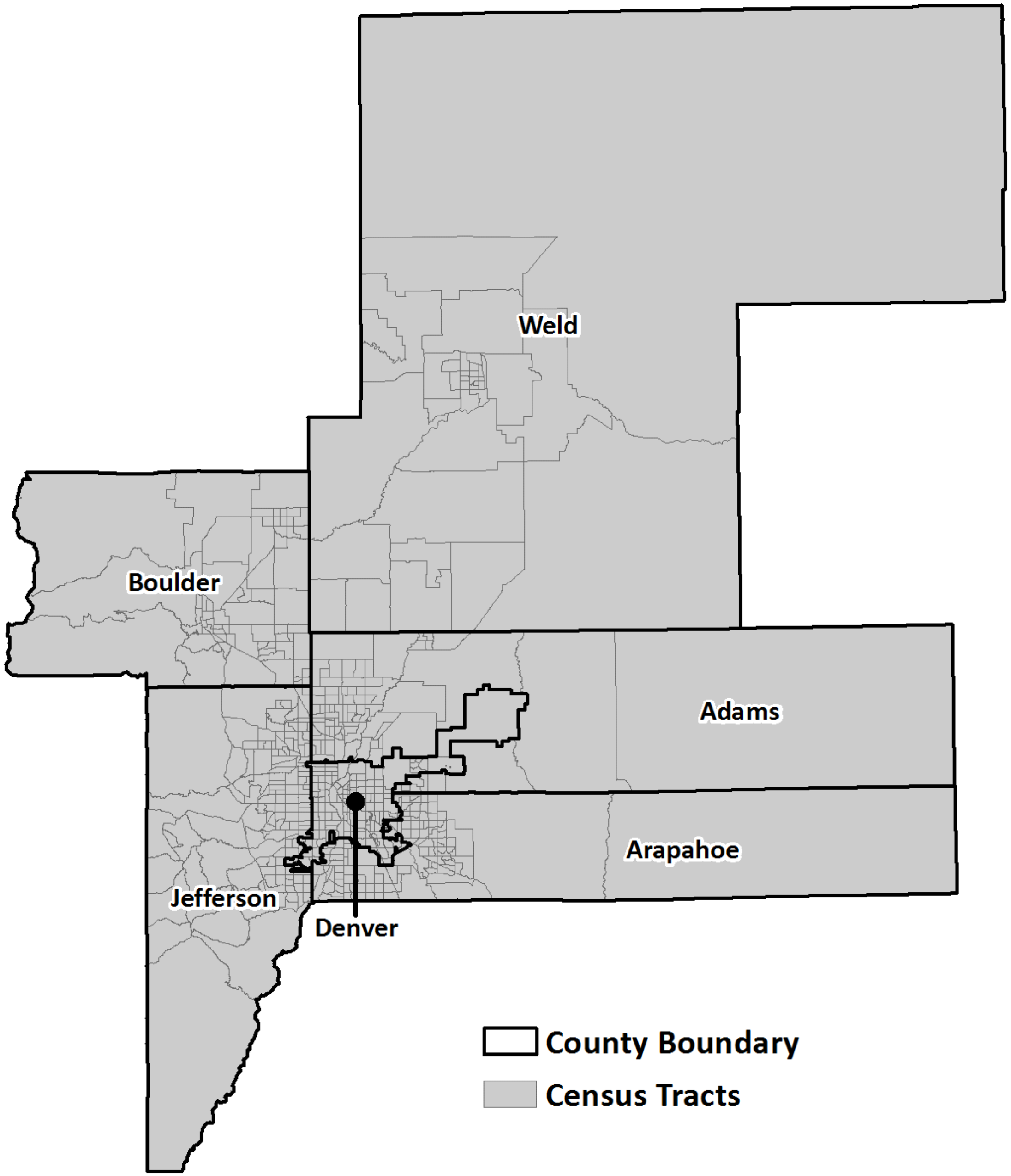
- Complementary topic in geography is dasymetric mapping *(Semenov-Tian-Shansky 1928; Wright 1936; Eicher & Brewer 2001; Mennis 2006; Riebel & Agrawal 2007)*

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

Deaths, All Causes (2000-2003)					
County	Total	% 75+	% Male	% Black	% Hisp
Adams	6,447	46	50	5	5
Arapahoe	8,378	56	48	8	8
Boulder	4,257	60	45	2	2
Denver	13,334	55	50	14	14
Jefferson	9,710	58	48	2	2
Weld	3,472	55	50	1	1

Male	Black	Hisp	Age	Census	Deaths	Synthetic
0	1	0	<35	53,999	181	53,818
0	1	0	35-44	22,597	263	22,334
0	1	0	45-54	22,128	492	21,636
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	1	65-74	204	1	203
1	0	1	75-84	74	4	70
1	0	1	85+	12	1	11

- 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

	M	B	H	A	D	Tract 1	Tract 2	Tract 3	...	Total
1	1	0	0	73	1	0.0055	0.0062	0.0078	...	1.0000
2	0	1	0	59	0	0.0055	0.0062	0.0078	...	1.0000
3	1	1	1	72	0	0.0055	0.0062	0.0078	...	1.0000
4	0	0	0	81	1	0.0055	0.0062	0.0078	...	1.0000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
N	0	0	1	35	0	0.0055	0.0062	0.0078	...	1.0000
					Total	2,850	3,228	4,047	...	516,902

	M	B	H	A	D	Tract 1	Tract 2	Tract 3	...	Total
1	1	1	0	73	1	0.0039	0.0060	0.0047	...	0.9779
2	0	1	0	59	0	0.0053	0.0070	0.0052	...	1.0041
3	1	1	1	72	0	0.0030	0.0068	0.0198	...	1.0647
4	0	0	0	81	1	0.0021	0.0027	0.0058	...	0.9025
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
N	0	0	1	35	0	0.0036	0.0054	0.0113	...	0.9825
					Total	2,850	3,228	4,047	...	516,902

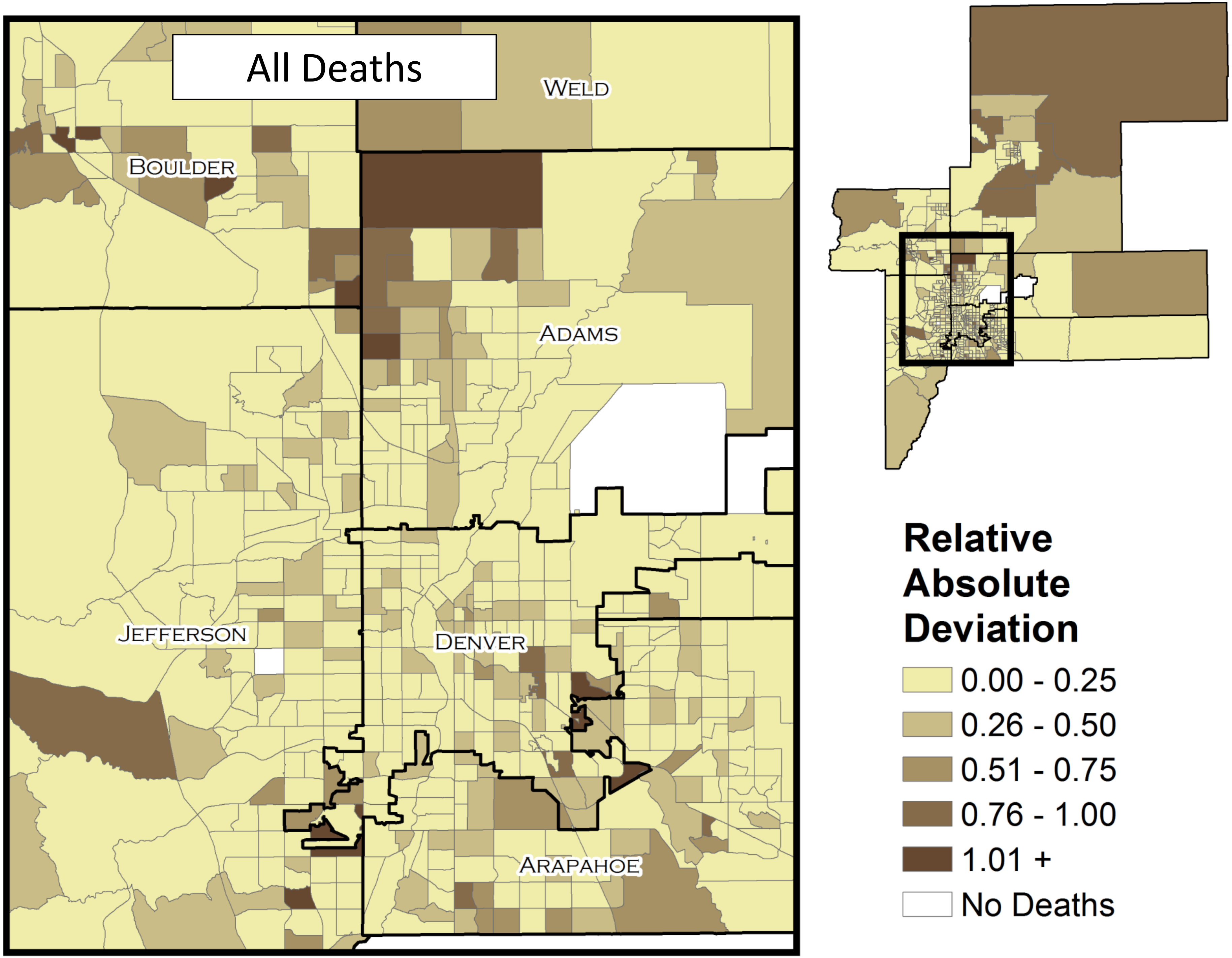
Validation

- 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

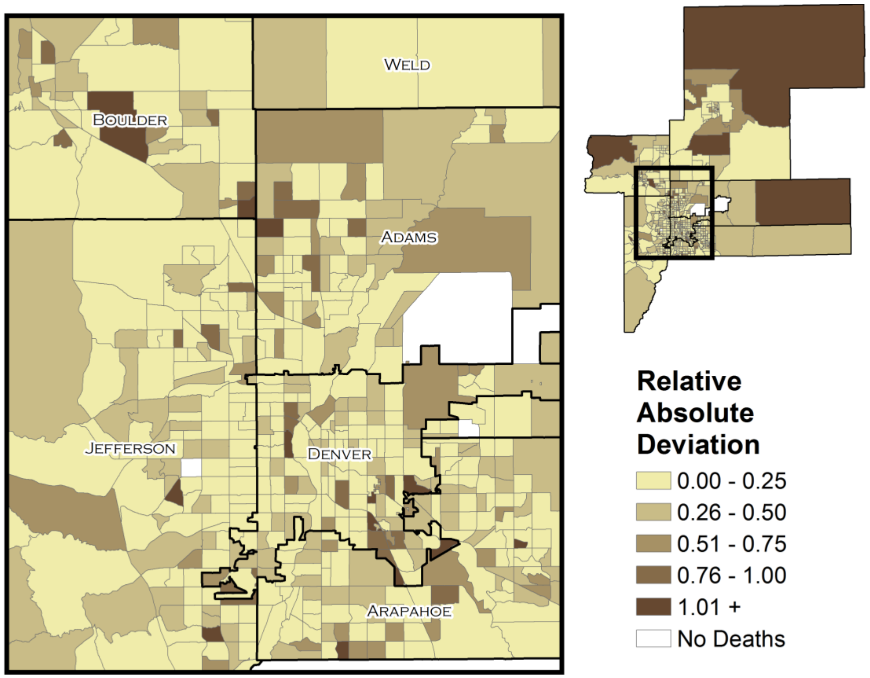
Denver County (135 tracts)						
Measure	All	Cancer	Heart	Stroke	Diabetes	Flu
Deaths	13,334	2,857	3,020	762	319	285
Spearman	0.86	0.81	0.82	0.67	0.50	0.56
MRAD	0.23	0.28	0.29	0.51	0.73	0.66

Total Metropolitan Area (576 tracts)						
Measure	All	Cancer	Heart	Stroke	Diabetes	Flu
Deaths	45,598	10,192	10,294	2,811	1,042	1,015
Spearman	0.90	0.84	0.86	0.74	0.49	0.61
MRAD	0.24	0.27	0.33	0.51	0.85	0.72

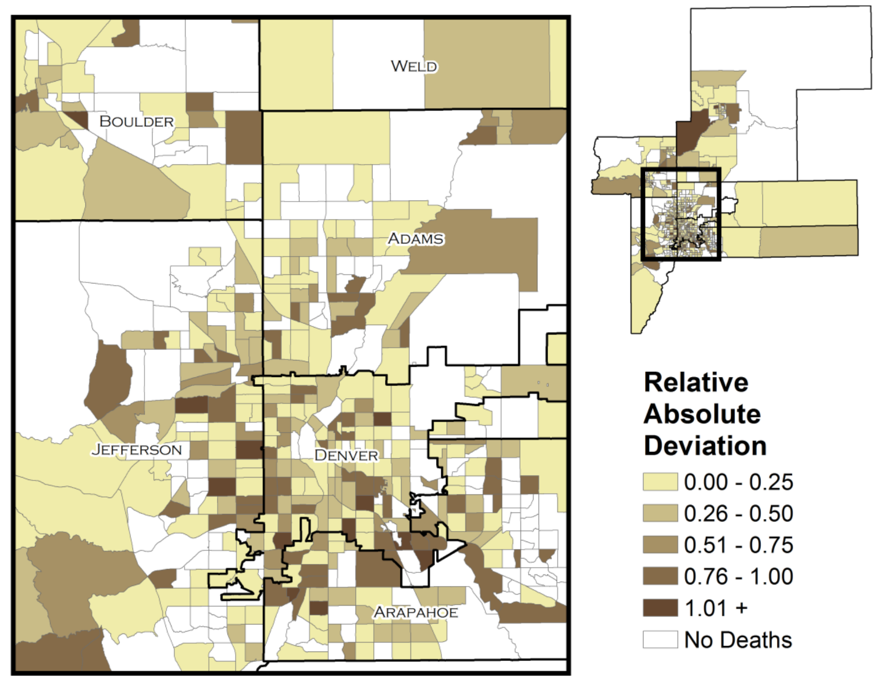

All Deaths
Weld
Boulder
Adams
Jefferson
Denver
Arapahoe
Relative Absolute Deviation
0.00 - 0.25
0.26 - 0.50
0.51 - 0.75
0.76 - 1.00
1.01 +
No Deaths

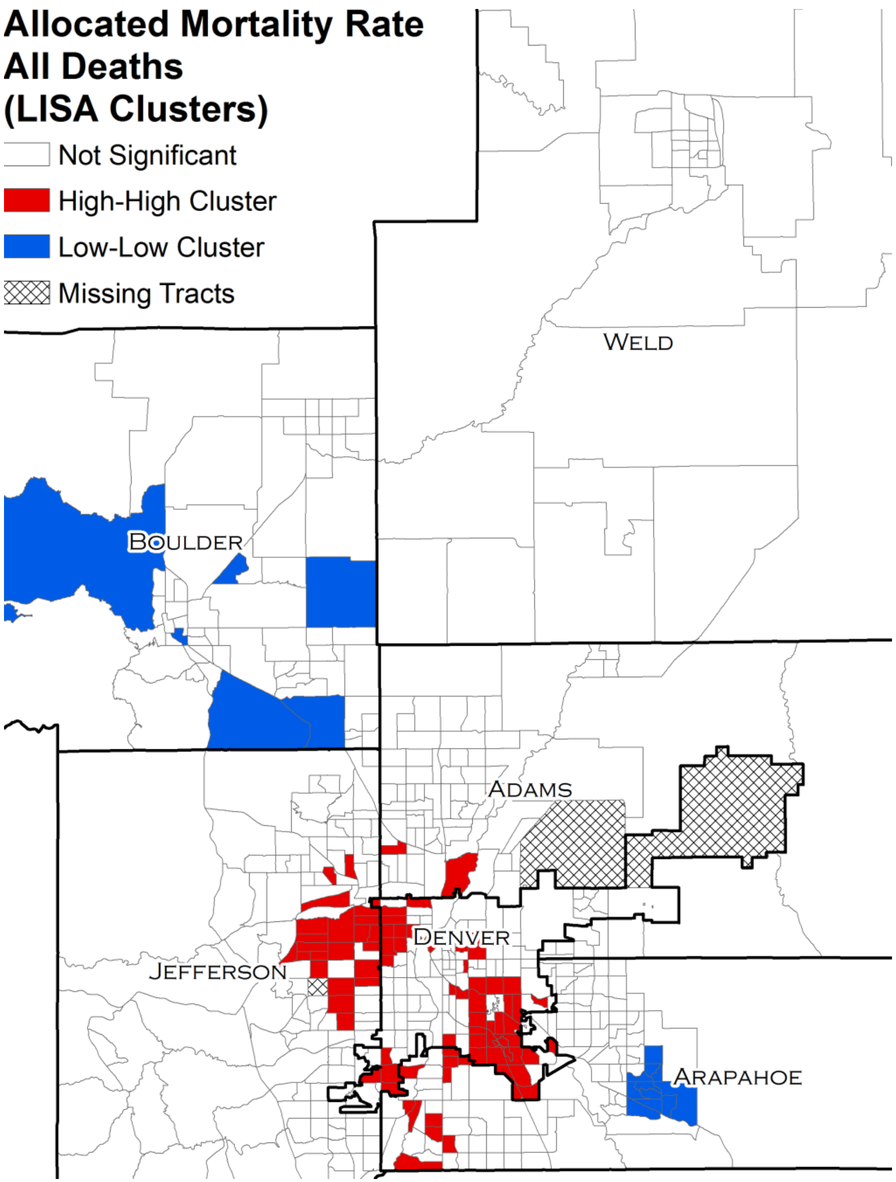
Validation Results (Cause-Specific)

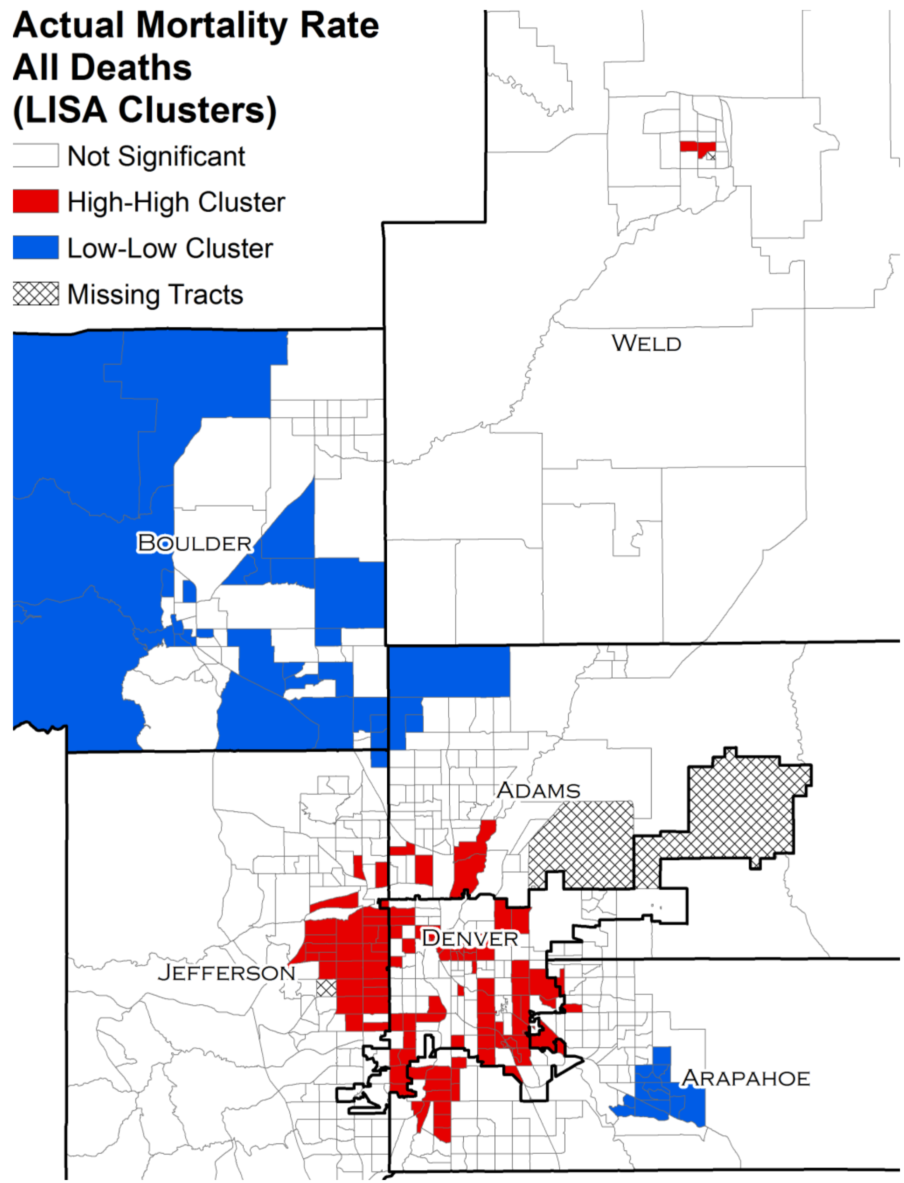
Cancer Deaths



Flu Deaths




Allocated Mortality Rate
All Deaths
(LISA Clusters)
Not Significant
High-High Cluster
Low-Low Cluster
Missing Tracts
Weld
Boulder
Adams
Denver
Jefferson
Arapahoe


Actual Mortality Rate
All Deaths
(LISA Clusters)
Not Significant
High-High Cluster
Low-Low Cluster
Missing Tracts
Weld
Boulder
Adams
Denver
Jefferson
Arapahoe

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

UNIVERSITY OF LOUISVILLE