Preprint

Accepted for publication in Annals of the Association of the American Geographers

Dasymetric modeling and uncertainty

Nicholas Nagle – University of Tennessee
Barbara Buttenfield – University of Colorado
Stefan Leyk – University of Colorado
Seth Spielman – University of Colorado

Abstract

Dasymetric models increase the spatial resolution of population data by incorporating related ancillary data layers. The role of uncertainty in dasymetric modeling has not been fully addressed as of yet. Uncertainty is usually present because most population data are themselves uncertain, and/or the geographic processes that connect population and the ancillary data layers are not precisely known. A new dasymetric methodology - the Penalized Maximum Entropy Dasymetric Model (P-MEDM) - is presented that enables these sources of uncertainty to be represented and modeled. The P-MEDM propagates uncertainty through the model and yields fine-resolution population estimates with associated measures of uncertainty. This methodology contains a number of other benefits of theoretical and practical interest. In dasymetric modeling, researchers often struggle with identifying a relationship between population and ancillary data layers. The PEDM model simplifies this step by unifying how ancillary data are included. The P-MEDM also allows a rich array of data to be included, with disparate spatial resolutions, attribute resolutions, and uncertainties. While the P-MEDM does not necessarily produce more precise estimates than do existing approaches, it does

help to unify how data enter the dasymetric model, it increases the types of data that may be used, and it allows geographers to characterize the quality of their dasymetric estimates. We present an application of the P-MEDM that includes household-level survey data combined with higher spatial resolution data such as from census tracts, block groups, and land cover classifications.

Key Words: dasymetric modeling, small area estimation, maximum entropy

Introduction

Researchers often desire population data at finer spatial resolutions than are publicly available. Fine resolution population data have been produced using such methods as areal interpolation, ordinary kriging, and dasymetric modeling. Dasymetric modeling produces fine-resolution estimates by utilizing relations between population distribution and ancillary geographic layers (Figure 1). Examples of ancillary data commonly used include: land cover (Mennis 2003), road density (Reibel and Bufalino 2005), slope (Schumacher et al. 2000), nighttime lights (Briggs et al. 2007), Landsat Thematic Mapper data (Yuan, Smith, and Limp 1997), LIDAR-derived building heights (Xie 2006; Kressler and Steinnocher 2008), IKONOS-derived land use classification (Liu, Kyriakidis, and Goodchild 2008), parcel data (Tapp 2010) and address points (Zandbergen 2011).

[Insert Figure 1 about here]

A common characteristic of these studies is the integration of: 1) population data, 2) ancillary data layers, and 3) a model relating the two. Each of these model inputs introduces its own uncertainty. For instance, population data may be the result of a sample survey rather than a complete enumeration, or they may represent the population distribution at a different point in time. Similarly, most geospatial ancillary layers contain their own sources of error in location or in attribution. Perhaps most important, however, are the uncertainties that arise from trying to quantitatively link or relate the population distribution to the ancillary layers. The link between the distribution of population and the distribution of other spatial attributes is typically uncertain and is difficult to quantify and assess.

Population distributions can never be completely characterized by deterministic relations because they are sensitive to local contextual factors such as cultural norms, local land use and housing regulations, local/regional environmental constraints, and the vagaries of random chance and history.

Despite these many sources of uncertainty, dasymetric techniques primarily treat this relational problem as a deterministic problem and do not adopt quantitative mechanisms for incorporating uncertainty. We propose a new methodology for dasymetric modeling that takes uncertainty into account, whether arising through uncertain population data, uncertain ancillary data, or uncertain relationships between population and ancillary data. Furthermore, this new methodology tracks these uncertainties through the dasymetric model in order to produce a measure of quality for the final population estimates.

This new model is based on an extension of Maximum Entropy (ME) techniques and we call it the Penalized Maximum Entropy Dasymetric Model (P-MEDM). ME techniques have been frequently used for "location-allocation" type problems in geography, in which an initial population needs to be allocated among a variety of different locations (Johnston and Pattie 1993). By themselves, ME techniques do not address the effects of uncertain input data layers, but we show how these uncertainties may be addressed by adding *penalties* to the ME. Furthermore, the P-MEDM model is shown to naturally adapt to the varying qualities of input data, allowing both high- and low-quality data to be utilized simultaneously.

The benefits of the P-MEDM include:

1. It integrates population and ancillary data sources with disparate levels of spatial and attribute resolution,
2. It incorporates information about the known uncertainty of the input population and ancillary data,
3. It incorporates information about uncertainty about the estimated relationship between population and ancillary data layers, and
4. It produces output population estimates with quantifiable uncertainty themselves, allowing other researchers to assess the quality of the final dasymetric products with regard to their own intended uses.

A simple case demonstrates the benefits of the P-MEDM. Since one of these benefits is its seamless ability to incorporate data of varying resolutions and qualities, this case study includes: household-level data from the Public Use Microdata Sample (PUMS) of the American Community Survey (ACS), tract- and block group-level summary data from the ACS, and land cover information from the National Land Cover Database (NLCD). The PUMS have extraordinary *demographic* resolution in that they contain records of individual households and the persons within them. They have very coarse *spatial* resolution, however. The enumeration districts for the PUMS (called Public Use Microdata Areas, or PUMAs) must contain at least 100,000 persons. While it is possible to identify in which PUMA a household lives, it is not possible to identify the household's location within the PUMA.

In contrast to the ACS microdata, the ACS census tract and block group data are aggregates of many households. While these layers have finer spatial resolution than the PUMS, they have coarser demographic resolution than the PUMS because they do not contain description of individual households or persons. The ACS is a rolling sample of the American population. On average, the ACS samples 135 household per census tract over a 5 year period, which are used to estimate the characteristics of the tract. Tract- and block group-level data layers are available for univariate population distributions, and some bivariate distributions, but they do not contain the richness of individual- and household-level data afforded by microdata. An additional problem when using ACS data for small areas is that these data often have large Margin of Errors. Because of the relatively small sample size of the ACS compared to the previous decennial census long form, tract and block group level estimates may be quite uncertain. The P-MEDM is able to account for this.

Finally, the case study includes relatively fine spatial resolution (30 meter) land cover data from the NLCD. These data, however, have relatively crude demographic characterization. Land cover data are valuable since they a proxy for demographic characteristics through a chain of indirect links that tie together land cover, land use, the type and density of housing, and the type and quantity of people residing in those houses. Thus, while land cover data have fine spatial resolution, they have

coarse demographic resolution. The P-MEDM model allows this type of proxy data to enter and play their own unique role within the dasymetric model.

This application case study demonstrates new types of analyses that are enabled by the P-MEDM. It also demonstrates how the uncertainty estimates produced by P-MEDM can provide signals to analysts indicating the quality of a dasymetric model and hence, to also indicate appropriate and inappropriate uses of dasymetric modeling more generally.

Areal Interpolation and Uncertainty

Areal interpolation is the process of downscaling coarse scale geographic data from *source regions* to a finer *target* scale (Goodchild and Lam 1980). Areal interpolation and dasymetric modeling are similar in that both are methods for downscaling spatial data. Areal interpolation differs from dasymetric modeling in that it assumes the boundaries of the target regions are arbitrary and that the variable being interpolated varies smoothly across the boundaries of the source and target regions. This smoothness assumption is explicit in methods such as Tobler's smooth pycnophylactic interpolation (1979) and Kyriakidis' area-to-area and area-to-point ordinary kriging (2005). In contrast, dasymetric modeling assumes that there are areas of relative homogeneity separated by sharp, discrete borders in the population surface, and dasymetric modeling seeks to incorporate ancillary data that are able to identify these boundaries. Uncertainty is reduced in dasymetric models if the boundaries between target regions correspond to real boundaries in the population surface. This assumption – whether the population surface is a smooth or a discrete surface – influences the nature of interpolation and the strategies chosen to address uncertainty (Goodchild and Lam 1980; Goodchild, Anselin and Deichmann 1995).

Assumptions about the nature of boundaries affect the kinds of ancillary information that are used for downscaling. The ancillary information in the smoothing methods includes information about the spatial structure (spatial

autocovariance) of the population surface; the ancillary information in dasymetric methods includes information that identifies boundaries of homogeneous regions the population surface. Stated differently, smoothing methods assume that the best ancillary information is information about the spatial autocorrelation of the population surface (i.e., the target variable itself), whereas dasymetric methods assume that the best information is information about the correlation between population and ancillary data at the same location. Both of these methods, however, involve interpolating the population to an unknown target region, and thus introduce uncertainty.

Many studies have been conducted to identify the sources of uncertainty in areal interpolation methods. In general, uncertainty is found to increase with larger source regions and smaller target regions (Sadahiro 2000). Ancillary data that can effectively define homogeneous regions also reduce uncertainty. Zandbergen and Ignizio (2010) compare different types of ancillary data and conclude that no one source of ancillary data is superior in all instances, but they do suggest that land cover data are relatively robust ancillary data for modeling population density. Furthermore, the variations between these methods can have significant impacts for subsequent spatial analysis. Maantay, Maroko and Hermann (2007) and Maantay, Maroko and Porter-Morgan (2008) used alternative interpolation techniques to estimate population density for an analysis of asthma prevalence, and showed that simpler areal interpolation estimators underestimated the negative effects compared to a more realistic, cadastral-based dasymetric model. These studies have indicated that the interpolation uncertainty is directly linked to the problem of identifying homogeneity in the population surface. Dasymetric models are successful if the ancillary data help to identify regions of homogeneous population. Similarly, the smoothing interpolation methods are successful if the population surface is smoothly varying with near-by areas being relatively homogeneous.

While these studies have provided a rich qualitative description of the sources of uncertainty, they have not provided a means to quantitatively assess the interpolation uncertainty within a specific study. The P-MEDM allows for the quantification of uncertainty in dasymetric estimates. Importantly, this method

quantifies not only the uncertainty that is inherent to the downscaling problem, but also the uncertainty that arises from using inexact ancillary data. While many studies have conducted validation experiments to compare the relative accuracy of different methodologies ex post facto, few methodologies allow the quantification of uncertainty as a direct model output. Notable exceptions are the geostatistical models, which do allow direct estimation of uncertainty (Kyriakidis 2005; Wu and Murray 2005; Liu, Kyriakidis and Goodchild 2008). These methods use the spatial autocorrelation of the population surface in order to quantify the uncertainty of downscaling, they do no, however, quantify the uncertainty that arises through using ancillary data of varying quality, nor do they provide an automatic means for balancing between these data of varying qualities. The P-MEDM has these capabilities.

METHODS

Dasymetric modeling includes a suite of techniques to more precisely depict the spatial distribution of population within the spatially aggregate regions (Slocum et al. 2009, chap. 15). Ancillary spatial data are essential to the dasymetric process. Dasymetric modelers often categorize ancillary data into two types: *limiting* and *related* ancillary variables. Limiting variables set constraints on the allowable population values, for example, by limiting population densities to zero in areas covered by water. Related ancillary variables can accommodate more complex relationships. For example, road density, elevation, or land cover might be used to amplify or constrain population densities. Objectively identifying these relations is a recurring problem, however. While there is no firmly established method for quantifying this relationship, linear regression techniques are common (Mennis 2009).

Intelligent Dasymetric Mapping

One of the most widely used and most flexible techniques for dasymetric modeling is *intelligent dasymetric mapping* (Flowerdew, Green, and Kehris 1991;

Mennis and Hultgren 2006). This technique downscales from source populations Pop_s to target populations $\widehat{\text{Pop}}_t$ as follows:

$$\widehat{\text{Pop}}_t = \text{Pop}_s \frac{w_t}{\sum_{t \in s} w_t} \quad (1)$$

where w_t is the expected population count in target region t. This expected population count w_t is derived through regression analysis using the population and ancillary layers; hence, Reibel and Agrawal (2007) call this approach *regression weighted dasymetric modeling.*

One practical benefit of this method is that the target populations are consistent with the source populations. Adding up the dasymetric estimates recovers the source populations: i.e. $\sum_{t \in s} \widehat{\text{Pop}}_t = \text{Pop}_s$. This constraint is frequently called the *pycnophylactic*, or volume-preserving, constraint (Tobler 1979).

Regression weighted dasymetric modeling does not account for uncertainty in knowing the inputs or their relationships. Often, the source population Pop_s is not precisely known. For example, the source population may come from a sample survey rather than a complete enumeration, or may represent the population at a different point in time. Thus, it might be more appropriate to write the source populations as $\widehat{\text{Pop}}_s$ in recognition of the fact that these data are actually estimates of populations whose size is unknown.

An implicit logic of equation (1) is that the target estimates $\widehat{\text{Pop}}_t$ are the result of rebalancing the target estimates w_t in order to be consistent with the source data.[1] We question, however, the logic of these constraints when the data are uncertain and noisy. Why should we exactly constrain our dasymetric results to inexact data? If these populations are derived through regression between the population data layers and the ancillary data layers (Mennis and Hultgren 2006; Reibel and Agrawal 2007), then these estimates contain uncertainty. Depending on the nature and quality of the data and on the strength of the regression relationship,

[1] Equation (1) is often presented in an alternative, but equivalent form: $\widehat{\text{Pop}}_t = w_t + \frac{w_t}{\sum_{t \in s} w_t}(\text{Pop}_s - \sum_{t \in s} w_t)$. In this form, the target estimates are seen as the result of allocating the source errors back to the target estimates w_t.

the uncertainty in these regression estimates may be quite large. Dasymetric modeling techniques do not currently address these uncertainties. Even more alarming, dasymetric models commonly produce a single estimate for the target population, implying that there is no uncertainty. There is no explicit recognition that estimation is uncertain, that estimation error is expected, nor of how large that estimation error might be. What is needed is a technique that can account for uncertainty, possibly balancing between the uncertainties in the source populations $\widehat{\text{Pop}}_s$ and the uncertainties in the target estimates w_t, in order to produce population estimates with quantifiable uncertainty. The P-MEDM model introduced here provides exactly these capabilities.

The Maximum Entropy approach to dasymetric modeling

We advance the maximum entropy (ME) framework as an effective tool for dasymetric modeling in presence of uncertainty. While ME has a long history in geography for location-allocation modeling (Wilson 1971), it is not widely used for areal interpolation (however, Mrozinski and Cromley 1999 is a noteworthy exception). The ME framework has been successfully employed, however, in the closely related problem of small area estimation, with applications varying from modeling the distribution of votes across subpopulations within electoral districts (Johnston and Pattie 1993) to estimating the distribution of subpopulations across census zones (Birkin and Clarke 1988; Wong 1992; Simpson and Tranmer 2005). Leyk et al studied the ME problem for small area estimation and assessed the degree of ambiguity in these subpopulation distributions (2013). These applications face problems that are essentially areal interpolation problems, although, they are not framed as such. A similar approach, developed independently, is the Bayesian Maximum Entropy Framework (Christakos 2000; Bogaert 2002).

In order to introduce maximum entropy modeling we present a stylized problem in the manner of Birkin and Clarke (1988). Table (1) presents a data situation in which there are two census tracts, with the goal to obtain the distribution of specific subpopulations -- as determined by race and income --

within these two tracts. There might be published data for the univariate distribution of race and univariate distribution of income by census tract, but the joint cross-tabulation of these attributes might not be published in order to protect anonymity of respondents. Thus, the modeling goal is to estimate these joint distributions.

Let w_{rck} be the population in a given row (r), column (c) and tract (k). The ME method estimates these missing values w_{rck} through solving the following problem:

$$\max \sum_{rck} \frac{w_{rck}}{d_{rck}} \log\left(\frac{w_{rck}}{d_{rck}}\right) \quad (2)$$

subject to the pycnophylactic constraints

$$\sum_r w_{rck} = \text{row sum, for each tract } k,$$

$$\sum_c w_{rck} = \text{column sum, for each tract } k,$$

and where d_{rck} are prior estimates. The prior estimates can be a constant value if no other information exists, or they can reflect prior, ancillary information about the joint distribution across attributes and/or location. For example, if the joint distribution of income and race is known for the overall population, then this overall distribution might be use as a prior estimate for each tract. This situation is displayed in Table 1. This situation was faced by Birkin and Clarke (1988) and Wong (1992), who obtained these global estimates from public use microdata. We adopt a similar strategy for integrating public use microdata into dasymetric modeling. Public use microdata are widely used across the social sciences, but are rarely encountered in geographic research. Thus, the potential impacts of this research are to provide techniques to "spatialize" household- and individual-level microdata and to provide new opportunities for interaction between geography and the other social sciences.

Penalized Maximum Entropy Dasymetric Modeling

A significant shortcoming of both the ME and regression weighted dasymetric modeling approaches is the failure to account for uncertainties. If the population data are uncertain, then the pycnophylactic constraints are too stringent. Models should not force population estimates to exactly add back up to other estimates that

are inherently imprecise. By relaxing the pycnophylactic constraints we incorporate uncertainty into the ME dasymetric approach.

Following Birkin and Clarke (1988) and Wong (1992), we begin with individual-level microdata, but we then place these data within the dasymetric modeling framework, and allocate these individuals to target regions according to small area census data as well as other ancillary spatial data layers. In survey sample data, each sampled individual represents a group of unsampled individuals in the population with the same characteristics. Let the target population w_{it} be the number of individuals like sample record i in target region t. Let $\widehat{\text{Pop}}_k$ be an uncertain estimate of a population group k, and let Pop_k be the true, but unknown population count. The index k is general; it could represent the total population in a tract, or in a target region, or a specific subpopulation, such as a tract level count of children in poverty, or even housing unit counts (not population counts) such as the count of single family housing units in a block group.

The ideal pycnophylactic constraints would be $\sum_{it \in k} w_{it} = \text{Pop}_k$. These constraints are not feasible, however, because we do not know the true values Pop_k. Thus, we must explicitly account for the unknown error between the true and estimated populations. We instead use the constraints: $\sum_{it \in k} w_{it} = \widehat{\text{Pop}}_k + e_k$, where e_k is the positive or negative error between the estimated and the true population count. These errors must be estimated in addition to the target populations w_{it}. Population estimates can come from a variety sources. Census estimates for various attributes and spatial resolutions can be used as one type of constraint. Other constraints might come from estimates produced by the researcher. For example, regression weighted dasymetric modeling requires the researcher to estimate a regression between population layers and ancillary land cover layers. Predictions from this regression could be another source of population estimates. This freedom to include many different sources of data into dasymetric modeling is a significant advancement. Any and all data layers that can be used to predict a population attribute become feasible inputs to dasymetric modeling.

These errors and constraints can be added to the ME model as follows: Choose w_{it} and e_k to

$$\max \sum_{it} \frac{n}{N} \frac{w_{it}}{d_{it}} \log\left(\frac{w_{it}}{d_{it}}\right) - \sum_{k} \frac{e_k^2}{2\sigma_k^2} \quad (3)$$

subject to the relaxed pycnophylactic constraints

$$\sum_{it \in k} w_{it} = \widehat{\text{Pop}}_k + e_k \text{ for each constraint } k,$$

where n is the size of the microdata sample, N is the population size, σ_k^2 is the variance of the uncertainty e_k, and d_{it} is a prior estimate of the population w_{it}. The term $\sum_k \frac{e_k^2}{2\sigma_k^2}$ is a penalty factor; it penalizes solutions with large errors e_k. If a P-MEDM solution exactly reproduces a population constraint i.e. $\sum_{it \in k} w_{it} = \widehat{\text{Pop}}_k$, then the error e_k will be zero, and there will be no penalty. However, if the solution does not exactly reproduce a population constraint, i.e. $\sum_{it \in k} w_{it} \neq \widehat{\text{Pop}}_k$, then there is an estimated error e_k, and the solution will be penalized for this discrepancy between the dasymetric estimates and the corresponding (ancillary) population estimates. This penalized Maximum Entropy framework, while new to geography, has received considerable treatment in the machine learning literature; see, for example, Chen and Rosenfeld (2000) and Dudik, Phillips and Schapire (2007).

The effect of the penalty term is to favor solutions with small errors e_k, and preferably with zero errors. Thus, the P-MEDM will tend to be close to the ancillary population estimates $\widehat{Pop}_k$. But not all population data are of equal quality, however. Some estimates are more reliable than others. The variance terms σ_k^2 in (3) accounts for the variation in data quality among the input data layers. If the variance σ^2 of a population estimate is relatively small then the penalty on errors e_k will be great. Thus, the P-MEDM solution will tend to have high fidelity to those input data that have a low variance. In contrast, if input data have a high variance, then the penalty effect will be small. Large errors e_k will be permissible for these imprecise population constraints. In the extreme case where the constraints are completely imprecise, then the penalty effect will be zero; it will be as if the constraint doesn't

exist. At the other extreme, when all the uncertainties are zero, then the P-MEDM solution is identical to the ME solution. A similar use of the model variance σ^2 as a qualitative measure for assessing the ancillary data was suggested by Mennis and Hultgren (2006); here, we have formalized and quantified this process.

The adaptive property is convenient because it allows many different types of ancillary data to be included. Data that are reliable will have a strong impact on the solution. Data that are not reliable will have a slight impact on the solution. If the solution can accommodate unreliable data without compromising the fit of reliable data, then it will do so. If, however, the P-MEDM cannot fit unreliable data without also compromising the fit of reliable data, then the reliable data take precedence.

The variance σ_k^2 must be known a priori, however, this is not usually a problem because the variance of population estimates is often available. For instance, if census tract- or block group-level summary tables are used as constraints, then the variance σ_k^2 is directly obtained from the Margins of Error that the Census Bureau publishes along with each estimate. Alternatively, if a population constraint is obtained by regression, such as a regression between population data and land cover data, then a prediction error variance σ_k^2 is provided by the regression model. In contrast to regression weighted dasymetric modeling, which ignores the fact that regression predictions are imprecise, or that the population data are themselves imprecise, P-MEDM is able to incorporate these uncertainties and to even adapt to variations in quality among the various input data layers.

A statistical motivation

In this section, we briefly describe a statistical motivation for the P-MEDM model, and describe techniques for estimating the uncertainty in the final dasymetric map product. The P-MEDM equation (3) has two parts; each of which is part of a specific, well-known log likelihood model. The first part - $\sum_{it} \frac{n}{N} \frac{w_{it}}{d_{it}} \log(\frac{w_{it}}{d_{it}})$, is proportional to the log-likelihood equation of a multinomial model for the weights w_{it} (Jaynes 2003). The second term in the P-MEDM - the sum

of square error terms $\frac{e_k^2}{2\sigma_k^2}$ - is proportional to the log likelihood of a Gaussian distribution. Together, these two terms represent the joint likelihood of the weights w_{it} and the errors e_k.

Taken together, we see that the P-MEDM equation is equivalent to simultaneously maximizing the likelihood of sample weights (which have a multinomial distribution) as well as the likelihood of the error distribution of ancillary population estimates (which are assumed to have a Gaussian distribution).

Since the P-MEDM problem is also a maximum likelihood problem, there are many possible ways to specify confidence intervals for the output dasymetric map. One way, based on the Likelihood Ratio, is conceptually simple, but computationally expensive. For the likelihood ratio method, one evaluates the P-MEDM model multiple times, once with the optimal w_{it} and errors e_k, and then repeatedly with alternative, sub-optimal values. Call the optimal log-likelihood value L_0, and the sub-optimal value L_1. The likelihood ratio is the ratio L_1/L_0. This ratio will have a χ^2 distribution (O'Brien 1992). For 95% confidence intervals of the weights, one can try different suboptimal weights w_{it}, repeatedly recalculating the likelihood ratio and searching for the weights that yield a likelihood ratio equal to the 2.5% and 97.5% values of a χ^2 distribution; this is a computationally expensive process.

Another, simpler method, relies on approximating the confidence intervals by a Gaussian distribution. For this method, we rely on the large sample approximation that the second derivatives of a likelihood function are inversely proportional to the covariance matrix of the model parameters (Schabenberger and Gotway 2004). With the covariance matrix in hand, one can then obtain the standard errors of the parameters, and the approximate 95% confidence intervals are obtained by the usual method of calculating $\pm$ 2standard errors. Many computational optimization procedures automatically calculate the second derivatives. Thus, this method is cheap to calculate; the P-MEDM only needs to be solved once and the final information about the second derivatives at the solution is immediately used to calculate the covariance matrix.

Case Study

Data

We demonstrate the P-MEDM model with a case study that models the population in Davidson County, Tennessee at various spatial and demographic resolutions. Davidson County contains the city of Nashville and had a population of about 600,000 persons in 2010. In order to demonstrate the wide applicability of the P-MEDM data, we incorporate three different sources of data representing four different levels of spatial resolution (Figure 2).

[Insert Figure 2 about here]

The first source of data are household-level microdata from the 5% Public Use Microdata Sample (PUMS) of the 2005-2009 American Community Survey. These microdata describe detailed characteristics of a 5% sample of individual households. The spatial resolution of the PUMS, however, is very coarse. The geography of the PUMS is a geographic unit called the Public Use Microdata Area (PUMA). PUMAs are large in order to preserve the anonymity of respondents; each PUMA contains at least 100,000 persons, and there are only 5 PUMAs in Davidson County. While these data are not widely used in geography, they are widely used in other social sciences. It is important that geographers develop tools that enable them to engage with the methods and findings being produced in the other social sciences; the P-MEDM enables the use of these data in spatial contexts.

The second source for data are summary tables from the 2005-2009 ACS representing two different geographic scales: census tracts (which are nested within PUMAs) and block groups (which are nested within tracts). These ACS summaries are survey *estimates* from a 10-12 percent survey of the population and are not actual population counts. For each ACS estimate, the Census Bureau also publishes a 90 percent Margin of Error (MOE), which is used to compute an error variance σ_k^2 (US Census Bureau 2009). The MOE can be quite large for small population

estimates, such as for some block groups and some small sub-populations. In extreme cases, the coefficient of variation for an ACS estimate may exceed 100 percent. Such a low precision makes the use of these data dubious in traditional dasymetric model. The P-MEDM, however naturally adapts to these differences in quality; these estimates can be added to the model with very little effect on the model. This reduces the burden on the researcher to subjectively evaluate which data are "good enough to use" and which are not.

The final source of data used in this case study is the 2006 National Land Cover Database (NLCD). The NLCD is raster grid with a spatial resolution of 30m covering the entire United States, with each cell preclassified into a single land cover type. Dasymetric models commonly use land cover data (Eicher and Brewer 2001; Mennis and Hultgren 2006; Reibel and Agrawal 2007; Tapp 2010). Despite their prevalence in dasymetric studies, land cover is not actually a direct estimate of any population attribute. In all such dasymetric studies, a relation between the land cover layer and population layer must be identified. We adapt the regression weighted areal interpolation approach (Reibel and Agrawal 2007; Mennis and Hultgren 2006) in order to derive ancillary population estimates $\widehat{Pop}$ and error variances σ^2 for each 30 meter pixel from the ancillary data.

Data processing and target zone construction

For this analysis, each land cover pixel was reclassified into six classes: five classes that are potential residential areas (these are used as related ancillary variables), and one non-residential class (this is used as a limiting ancillary variable). The five land cover classes used as related ancillary variables are High-, Medium- and Low-Intensity and Open Space, Developed land (NLCD classes 24, 23, 22 and 21), and Vegetated land (all pixels not classified as water, barren or wetland). The underlying NLCD definitions for these classes indicate that residential land use is possible (Fry et al. 2011). Water, barren land and wetland are grouped into a single non-residential class since the given class definitions indicate that residential land use is highly unlikely (we ignore the possibility, for example, of residential houseboats). We then merged the NLCD pixels within census block

groups in order to construct target regions. Thus, target regions for this analysis are sub-block group regions with homogeneous NLCD classification; there are at most six target regions within each block group.

In order to construct population constraints, the ACS data are processed by selecting a subset of the summary tables and then further collapsing some categories. The census tract and block group data tables that are used as ancillary estimates are

- Total population and Number of housing units.
- Number of housing units by building type (six categories: single family detached, single family attached, a building with 2-9 units, one with 10-49 units, one with 50 units or more, and housing units not elsewhere classified).
- Number of households by tenure status (two categories: Own or Rent).
- Number of households by household Income (in three categories: <=$25,000, $25,001-50,000, and >=$50,001).
- Number of households by race of householder (two categories: Black, All Other).
- Number of households by income and race (for a total of six categories).
- Number of households by income and tenure (six categories).
- Number of households by race and tenure (four categories).

These data constraints and their geographic scale are summarized in Table 2.

[Insert Table 2 about here.]

The Census Bureau publishes each of these tables at both the block group and tract level, except for the Households by Income and Tenure and the Households by Income and Race tables, which are only published down to the tract level. While it might seem that the constraints at both the tract and the block group levels are redundant, this is not the case. Even though the tract level estimates are equal to the sum of block group estimates, the MOEs can be very different. The MOE of the sum of block group estimates is at least as large as the MOE of the tract level estimate. Thus, including the tract level estimates in addition to those for the block groups allows the P-MEDM to maintain more fidelity to the tract level estimates.

The tract and block group summaries are already in the proper form that allows us to use them as pycnophylactic constraints with error, i.e. $\mathrm{Pop}_k = \widehat{\mathrm{Pop}_k}^{\mathrm{ACS}} + e_k$.

The related NLCD land cover data are not immediately usable as population constraints, however; we must convert the land cover data into constraining population estimates. Following the regression weighted dasymetric modeling approach, we use regression techniques to specify a relationship between the land cover and population layers and then calculate the associated standard errors of prediction.

Many dasymetric studies have specified a linear regression relation between land cover and population density. Logically, however, land cover is less related to population density than it is to building type and building density. While this type of relation was difficult to specify and use in previous dasymetric studies, it can be easily used in the P-MEDM approach. We use a Poisson Generalized Linear Model to relate housing unit counts to land cover type. Let $\mathrm{HU}_{cb}^{(k)}$ denote the number of housing units of building type k in block group b and land cover c. We assume that this quantity has a Poisson distribution. We then specify the following link between expected number of housing units and the land cover:

$$E(\mathrm{HU}_{b}^{(k)}) = \sum_{c} \mathrm{Area}_{cb}\beta_{c}^{(k)}$$

where $E(\mathrm{HU}_{b}^{(k)})$ is the expected number of housing units, Area_{cb} is the land area of block group b with land cover c, and $\beta_{c}^{(k)}$ is the regression coefficient measuring the housing unit density for building type k in land cover class c. This characterizes a Poisson Generalized Linear Model with additive link function (McCullagh and Nelder 1989). We have also added overdispersion (extra variance) to the model specification. Overdispersion may arise from either uncertainty in measuring the dependent variable, which is certainty present for these data measured by the ACS, or from model misspecification, or from measurement error in the land cover classification, which is certainly present as well.

Once this regression is fitted for each building type k, we construct estimates $\widehat{\mathrm{HU}}^{(k)}$ for each target region. Thus, the regression produces a geographic data layer containing housing unit predictions for each building type and target region. In addition to producing a regression prediction, the regression also produces a

prediction variance; this is the σ^2 that is needed for P-MEDM. Thus, we use regression to construct housing unit estimates for each target region, which are then added to the P-MEDM constraints, along with the constraining tract and block group estimates that are provided directly by the Census Bureau.

An important note to make is that, while we have explicitly quantified the prediction accuracy of the land cover data, we have not explicitly quantified the classification error of the NLCD. The NLCD is itself a complex dataset, and there is an extensive literature on uncertainty and error in NLCD classifications (Wickham et al 2010; 2013). This classification error is especially problematic in rural areas, where isolated housing units are often misclassified as Vegetated. This type of error is not explicitly incorporated in the P-MEDM. It is implicitly incorporated, however, as the misclassification error will reduce the prediction accuracy of the land cover data, and thus increase the prediction variance σ^2 of these P-MEDM constraints. More accurate land cover data would presumably lead to higher prediction accuracies. This misclassification problem is a common impediment to dasymetric modeling, is subject to intensive research in the land cover/ remote sensing community and can only be solved by improved detection procedures.

Results

In this section, we present a variety of different products and analyses that are made possible by the P-MEDM technique. Once weights w_{it} are obtained, they can be utilized in a variety of different ways.

Dasymetric mapping

First, it is possible to duplicate traditional dasymetric modeling. The sum $\sum_i \ w_{it}$ will produce population estimates for each target region. Since we have the microdata, with all of their individual- and household-level attributes, it is possible to obtain dasymetric maps for various subpopulations as well. Let k be any subpopulation of interest, then $\sum_{i \in k} w_{it}$ represents the subpopulation estimate for

target region t. Figure 3 displays both a dasymetric map for total population, as well as low income black households, in Davidson County.

[Insert Figure 3 about here]

The P-MEDM makes it possible, with one optimization, to produce many different types of dasymetric estimates at the target zone scale. This contrasts with existing practice, in which the dasymetric model must be independently fitted for each estimate. Eicher and Brewer (2001), for example fit dasymetric models separately for total population, Hispanic persons, and number of children. The P-MEDM approach can provide these different estimates simultaneously. Furthermore, the P-MEDM estimates are internally consistent in the sense that dasymetric estimates for sub-populations will add up to the dasymetric estimates for larger populations.

Small Area estimation

The P-MEDM also allows the creation of new small area estimates for use in mapping and analysis. Consider, for example, an analysis of racial disparities in rates of homeownership across neighborhoods. Ideally, such an analysis, would control for the confounding effects of income. In order to do this, we need neighborhood-level estimates of the trivariate table containing tenure (own versus rent), race, and income. This trivariate table is not produced by the Census Bureau. The Census produces each of the bivariate tables for census tracts, but not the trivariate table. Researchers might use the PUMS to construct the trivariate table for each PUMA, but PUMAs are so large that they are inadequate proxies to assess any neighborhood effects.

This type of estimate is possible with the P-MEDM approach. Using the estimated weights w_{it}, we can construct the necessary trivariate tables for each tract (or block group, or target region) and then directly calculate the homeownership rate for each region. Figure 4 displays maps of the estimated home ownership rates for each tract, separated by race of the householder and household income. These

maps are not possible from the ACS summary tables. Similarly, if the PUMS were used to estimate these homeownership rates, these rates would have to be constant across all tracts within the same PUMA. The P-MEDM clearly shows that there is likely spatial variation in these homeownership rates across the county. In this way, the P-MEDM technique allows richer mapping and analytical capabilities than currently exist.

[Insert Figure 4 about here]

Uncertainty analysis

Finally, we demonstrate the ability of the P-MEDM to produce estimates of uncertainty for the output maps and estimates. For the homeownership analysis in the previous section, we might consider the odds ratio of homeownership between black and white households, for households with incomes between $25,001 and $50,000. Using the estimated data that went into Figure 4, we could produce the single best estimate of the odds ratio for each tract. But for scientifically robust analysis, it is necessary to evaluate the uncertainty of these estimates as well.

This is possible with the P-MEDM approach. Using the second derivatives of the likelihood equation and the covariance function, we have simulated 100 different sets of weights w_{it}^{sim}. For each simulation, we aggregate the weights and calculate the odds ratio for each simulation. These simulations give us a Monte Carlo estimate of the statistical uncertainty (see, for example, Wood (2006, pp. 246-7) for discussion of this technique in the context of penalized generalized linear models).

Figure 5 displays box plots of the simulated odds ratio for each census tract. Each of these estimates is in some sense consistent with the input data, subject to the inherent uncertainties in both input data, as well as in the maximum entropy/maximum likelihood estimating procedure. Each column of the figure represents a different census tract, with the tracts sorted from that tract with the highest proportion of white households on the left to that tract with the highest proportion of black households on the right. For households earning $25,001 to

$50,000 , we see clearly that black households are more likely to own a home in predominantly black neighborhoods than are white households, and that this is the opposite in predominantly white neighborhoods. Also, in a majority of tracts, black households are less likely to own a home than are white households (more tracts have a an odds ratio less than 1.0 than have an odds ratio above 1.0). Even considering the error bars, this trend is evident and robust.

[Insert Figure 5 about here]

It is important to emphasize that not all tracts are reliably estimated. The error bars can be quite large. Taken individually, many of the tract-level estimates might have been deemed unreliable or unusable. A choropleth plot, for instance, might not have robust class breaks. This uncertainty analysis, however, allows researchers to make this determination on their own. This is not possible with current dasymetric techniques. Despite the high uncertainty of individual estimates, we can still see from Figure 5 that robust scientific generalizations are possible. Even with many census tracts imprecisely estimated, there is a clearly identified relationship between the odds ratio of homeownership and the racial composition of a census tract. This type of uncertainty information is important since it allows researchers to objectively evaluate the quality of dasymetric estimates. This evaluation of quality will vary by researcher and by data use. With the P-MEDM estimate, researchers are provided with sufficient information to undertake such an evaluation, even if they know very little about dasymetric or small area estimation techniques.

Discussion

This article has introduced a new methodology for dasymetric modeling. This model is able to account for information about the quality of ancillary data with regard to downscaling population estimates. Previous literature has focused on *spatial* accuracy, noting that finer resolution population data are preferred to coarse resolution data. The results here suggest a cautionary note that does not exist in the

previous literature; data with finer spatial resolution are often less accurate. For example, block groups are less accurate than tracts. Thus, while we expected the tract level data to be redundant given the block group data, this was not the case; the tract level data were more precise than the sum of the block group data. Similarly, data with even higher resolution, such as land cover have even less accurate information about population. Future developments in dasymetric modeling should consider more fully the multidimensional nature of tradeoffs between different data sources and more fully acknowledge tradeoffs between spatial resolution and data accuracy.

The P-MEDM model does have some potential limitations. First and foremost, the model requires an estimate of variance for each ancillary variable. This may not be as challenging as it first seems, however. For example, census data in the United States are published with estimates of the Margin of Error. Secondly, ancillary data are important only insofar as they are able to predict population well. For example, the quality of the land cover class is only indirectly important, what is directly relevant is how well the available land cover data can predict the population surface. This can be determined by regression techniques, regardless of the quality of the land cover data. More precise land cover data will have higher predictive power, but it is the predictive power that is the directly relevant measure of data quality, not the quality of the land cover classification. Whatever this predictive fit between the ancillary data and the population distribution is, the P-MEDM will properly find the balance between the different ancillary data inputs. It is possible that there are situations in which it is still difficult to quantify the uncertainty of a particular ancillary data layer, but we believe that the framework described here is general enough to incorporate many different types of data.

A second limitation of the P-MEDM is that it does not incorporate spatial autocorrelation of population data as ancillary component such as in smoothing techniques that are motivated by the concept of areal interpolation. Integrating smoothing and the described P-MEDM is difficult because traditional statistical smoothing methods rely on a Gaussian assumption, which is incompatible with the log-linear assumption used here. The log-linear assumption is convenient because it

is guaranteed to produce non-negative population estimates. This is not true, for example, with common geostatistical techniques, where ad hoc fixes are sometimes needed to enforce non-negativity (Yoo and Kyriakidis 2006). Additionally, geostatistical models are tailored to specific sub-populations; there is no guarantee, for instance, that the spatial structure of low-income Black households is the same as the spatial structure of other household types. The P-MEDM approach is attractive because it can model all subpopulations found in the microdata simultaneously. Nonetheless, explicit incorporation of spatial autocorrelation should enhance the predictive ability of the P-MEDM, and further research will investigate this possibility.

Another limitation is that we have not used explicitly spatial regression models in the modeling of ancillary data. We have assumed that the errors in the constraint equations are uncorrelated. This, however, is a limitation of our implementation and not of the P-MEDM. One way to incorporate spatial autocorrelation among the constraints would be to change the penalties to include the entire inverse covariance matrix of the constraints. In essence, rather than assuming that the penalties derive from an approximating Gaussian distribution, this would instead treat the errors as if derived from a multivariate Gaussian distribution. At an intuitive level, this modification would decluster, or decorrelate, the constraints, giving more weight to the constraints that are precise or that are relatively uncorrelated with other constraints. For ancillary data that are included through a regression relationship, as the land cover data are here, it would be possible to use a linear or nonlinear spatial regression model, and account for spatial autocorrelation explicitly. Incorporating the spatial autocorrelation of small area census data will be more difficult, however, as the spatial structure of survey sampling errors is relatively understudied. These are subjects for further research.

Summary

We have developed a new conceptual framework for dasymetric modeling called the Penalized Maximum Entropy Dasymetric Model (P-MEDM). This P-MEDM

addresses four problems that have challenged recent dasymetric modeling approaches, those of:

1. Accounting for uncertainty in the dasymetric output,
2. Accounting for uncertainty in the relationship between ancillary variables and the target variables,
3. Accounting for uncertainty in the population data themselves, and
4. Simultaneously producing estimates for multiple subpopulations.

The P-MEDM technique is able to integrate data with disparate levels of spatial and demographic resolution in order to construct richer and more complete population models at finer spatial scales. The penalizing mechanism allows the P-MEDM to adjust automatically to various input data having different levels of precision. This property will allow future dasymetric modeling efforts to consider multiple ancillary data sources, regardless of their quality. This reduces the need for modelers to subjectively evaluate which ancillary data are "good enough" and which are not. The P-MEDM model is also able to quantify the uncertainty of the final product, which has been largely ignored to date. This is an important factor in making dasymetric techniques accessible to other social scientists and to support new applications in other disciplines. The quantification of statistical uncertainty will make it possible for potential users to objectively evaluate the output of the dasymetric model, even if they do not fully understand the dasymetric model itself. Researchers regularly use data for which they do not fully understand the estimation procedure, but in order for them to evaluate their analysis, it is crucial that they are given the sufficient information to determine the quality of these data.

As demonstrated in the case study with the P-MEDM approach, it is possible to produce new dasymetric data or modeling tools that are usable by the general social science research community. Survey weighting, as used in this research, makes it possible for users to effectively analyze a wide variety of attributes. This article proposes dasymetric modeling as one effective strategy to producing general purpose spatial microdata, acceptable for use in a wide variety of research applications.

All computations in this article were produced in the R statistical computing environment, and programs and data are available from the corresponding author upon request.

References

Birkin, M., and M. Clarke. 1988. "SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples." *Environment and Planning A* 20: 1645–1671.

Briggs, D. J., J. Gulliver, D. Fecht, and D. M. Vienneau. 2007. "Dasymetric modelling of small-area population distribution using land cover and light emissions data." *Remote Sensing of Environment* 108: 451–466.

Bogaert, P. 2002. "Spatial prediction of categorical variables: the Bayesian maximum entropy approach." *Stochastic Environmental Research and Risk Assessment.* 16: 425-448.

Chen, S.F. and R. Rosenfeld. 2000. "A survey of smoothing techniques for ME models." *IEEE Transcactions on Speech and Audio Processing.* 8(1): 37-50.

Christakos, G. 2000. *Modern Spatiotemporal Geostatistics.* Oxford University Press, New York.

Dudik, M., S. F. Phillips, R. E. Schapire. 2007. "Maximum entropy density estimation with generalized regularization and an application to species distribution modeling." *Journal of Machine Learning Research*. 8: 1217-1260.

Eicher, C. L., and C. A. Brewer. 2001. "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation." *Cartography and Geographic Information Science* 28: 125–138.

Flowerdew, R., M. Green, and E. Kehris. 1991. "Using areal interpolation methods in geographic information systems." *Papers in Regional Science* 70: 303–315.

Fry, J., G. Xian, S. Jin, J. Dewitz, C. Homer, L. Yang, C. Barnes, N. Herold, and J. Wickham. 2011. "Completion of the 2006 National Land Cover Database for the Conterminous United States" *Photogrammetric Engineering and Remote Sensing* 77(9), pp. 858–864.

Goodchild, M.F., L. Anselin and C. Deichmann. 1995. "A framework for the areal interpolation of socio-economic data." *Environment and Planning A* 25: 383-397.

Goodchild, M. F. and N. Lam. 1980. "Areal interpolation: A variant of the traditional sptaial problem." *Geo-Processing*. 1: 297-312.

Janes, E. T. 2003. *Probability Theory: The Logic of Science.* Cambridge University Press, London.

Johnston, R. J., and C. J. Pattie. 1993. "Entropy-Maximizing and the Iterative Proportional Fitting Procedure." *The Professional Geographer* 45: 317–322.

Kressler, F., and K. Steinnocher. 2008. "Object-oriented analysis of image and LiDAR data and its potential for a dasymetric mapping application." In *Object-Based Image Analysis*, ed. Thomas Blaschke, Stefan Lang, and Geoffrey J. Hay, 611–624. Springer Berlin Heidelberg.

Kyriakidis, P. C. 2004. "A geostatistical framework for area-to-point spatial interpolation." *Geographical Analysis* 36 (July): 259–289.

Leyk, S., B. P. Buttenfield, and N. N. Nagle. 2013. "Modeling ambiguity in census microdata allocations to improve demographic small area estimates." *Transactions in Geographic Information Science*. Published online: 7 January, 2013.

Liu, X. H., P. C. Kyriakidis, and M. F. Goodchild. 2008. "Population-density estimation using regression and area-to-point residual kriging." *International Journal of Geographical Information Science* 22 (mar): 431–447.

Maantay, J. A., A. R. Maroko and C. Herrmann. 2007. "Mapping population distribution in the urban environment: The Cadastral-based Expert Dasymetric System (CEDS)." *Cartography and Geographic Information Science* 34(2): 77-102.

Maantay, J. A., A. R. Maroko, and H. Porter-Morgan. 2008. "Research note—A new method for mapping population and understanding the spatial dynamics of disease in urban areas: Asthma in the Bronx, New York." *Urban Geography* 29(7): 724-738.

McCullagh, P. and J. Nelder. 1989. *Generalized Linear Models*. Chapman and Hall/CRC Press.

Mennis, J. 2003. "Generating Surface Models of Population Using Dasymetric Mapping." *The Professional Geographer* 55: 31–42.

———. 2009. "Dasymetric Mapping for Estimating Population in Small Areas." *Geography Compass* 3: 727–745.

Mennis, J., and T. Hultgren. 2006. “Intelligent Dasymetric Mapping and Its Application to Areal Interpolation.” *Cartography and Geographic Information Science* 33: 179–194.

Mrozinski, R. D., and R. G. Cromley. 1999. “Singly- and Doubly-Constrained Methods of Areal Interpolation for Vector-based GIS.” *Transactions in GIS* 3: 285–301.

O’Brien, L. 1992. *Introducing Quantitative Geography: Measurement, Methods and Generalized Linear Models*. Routledge.

Reibel, M., and M. E. Bufalino. 2005. “Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems.” *Environment and Planning A* 27: 127–139.

Reibel, M., and D. Agrawal. 2007. “Areal Interpolation of Population Counts Using Pre-classified Land Cover Data.” *Population Research and Policy Review* 26 (5): 619–633.

Sadahiro, Y. 2000. “Accuracy of count data transferred through the areal weighting interpolation method.” *International Journal of Geographical Information Science*, 14(1): 25-50.

Schabenberger, O. and C. A. Gotway. 2005. *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC Press.

Schumacher, J. V., R. L. Redmond, M. M. Hart, and M. E. Jensen. 2000. “Mapping Patterns of Human Use and Potential Resource Conflicts on Public Lands.” *Environmental Monitoring and Assessment* 64 (1): 127–137.

Simpson, L., and M. Tranmer. 2005. “Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software.” *The Professional Geographer* 57: 222–234.

Slocum, T. A., R. B. McMaster, F. C. Kessler, and H. H. Howard. 2009. *Thematic Cartography and Geovisualization*. Pearson Prentice Hall.

Tapp, A. F. 2010. “Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources.” *Cartography and Geographic Information Science* 37: 215–228.

Tobler, W. R. 1979. "Smooth Pycnophylactic Interpolation for Geographical Regions." *Journal of the American Statistical Association* 74 (September): 519–530.

U. S. Census Bureau. 2009. "A Compass for Understanding and Using American Community Survey Data: What Researchers Need to Know."

Wickham, J.D. S.V. Stehman, J.A. Fry, J.H. Smith, C.G. Homer. (2010). "Thematic accuracy of the NLCD 2001 land cover for the coterminous United States." *Remote Sensing of the Environment* 114: 1286-1296.

Wickham, J. D. Stehman, S.V., Gass, L., Dewitz, J., Fry, J. A., Wade, T. G. (2013) Accuracy assessment of NLCD 2006 land cover and impervious surface. *Remote Sensing of Environment* 130, 15, pp. 294-304.

Wilson, A. G. 1971. "A family of spatial interaction models, and associated developments." *Environment and Planning A* 3: 1–32.

Wong, D. W. S. 1992. "The Reliability of Using the Iterative Proportional Fitting Procedures." *The Professional Geographer* 44: 340–348.

Wood, S. 2006. *Generalized Additive Models: An introduction with R.* Chapman & Hall/CRC Press.

Wu, C and A. T. Murray. 2005. "A cokriging method for estimating population density in urban areas." *Computers, Environment and Urban Systems*. 29: 558-579.

Xie, Z. 2006. "A framework for interpolating the population surface at residential-housing-unit level." *GIScience and Remote Sensing*. 43(3): 1-19.

Yoo, E.-H. and P.C. Kyriakidis. 2006. "Area-to-point kriging with inequality-type data." *Journal of Geographical Systems*. 8: 357-390.

Yuan, Y., R. M. Smith, and W. F. Limp. 1997. "Remodeling census population with spatial information from LandSat TM imagery." *Computers, Environment and Urban Systems* 21: 245–258.

Zandbergen, P. A. 2011. "Dasymetric Mapping Using High Resolution Address Point Datasets." *Transactions in GIS* 15: 5–27.

Zandbergen, P. A. and D. A. Ignizio. 2011. "Comparison of dasymetric mapping techniques for small-area population estimates." *Cartography and Geographic Information Science*, 37(3): 199-214.

Table Captions

Table 1. A simplified scenario involving tract-level summary data and region-wide joint distributions. Missing cell values are denoted by a question mark. Maximum Entropy can estimate these missing values.

Table 2. Listing of calibration constraints by geographic detail and data source

Table 1

	Tract 1			Tract 2			Region Total		
	Own	Rent	Total	Own	Rent	Total	Own	Rent	Total
Black	?	?	5	?	?	10	10	5	15
White	?	?	35	?	?	25	10	50	60
Total	10	30	40	10	25	35	20	55	75

1 Table 2

2

Constraint	Tract	Block Group	NLCD Region	Source
Total Pop	X	X		ACS
Housing Units	X	X		ACS
Income	X	X		ACS
Tenure	X	X		ACS
Race	X	X		ACS
Tenure X Race	X	X		ACS
Income X Tenure	X			ACS
Income X Race	X			ACS
Building Type	X	X		ACS
Building Type			X	Regression(ACS and NLCD)

3

4

Figure Captions

Figure 1. Dasymetric models utilize ancillary data in order to produce population estimates at finer resolution. As shown here, coarse resolution population data are combined with a gridded land cover layer in order to produce a gridded estimate of population. This article presents a method to track uncertainty from the input layers to the output map.

Figure 2. Davidson county illustrated using data at different spatial scales. Davidson county has 5 PUMAs (thick black line), 144 census tracts (thin black lines) and 467 block groups (grey lines). The base layer is a hill-shaded representation of the 2006 National Land Cover Database, residential classes in red shades, and vegetated classes in green.

Figure 3. Dasymetric maps for total population (right) and for persons living in a home with a Black householder and household income less than $25,001 (left). These dasymetric estimates were produced using the exact same model, as described in the text. Other population characteristics can be similarly modeled and mapped.

Figure 4. Maps of tract-level homeownership rates, separated by race of householder (varying across columns) and household income (varying across rows). These estimates are not published by the Census Bureau and have been estimated according to the P-MEDM as described in the text.

Figure 5. Boxplots of simulated odds ratios for each census tract. The odds ratios are those of homeownership for black households relative to white households, each with household incomes between $25,001 and $50,000. Odds ratios below 1.0 indicate that black households are less likely to own their home than are white households.

Figure 1

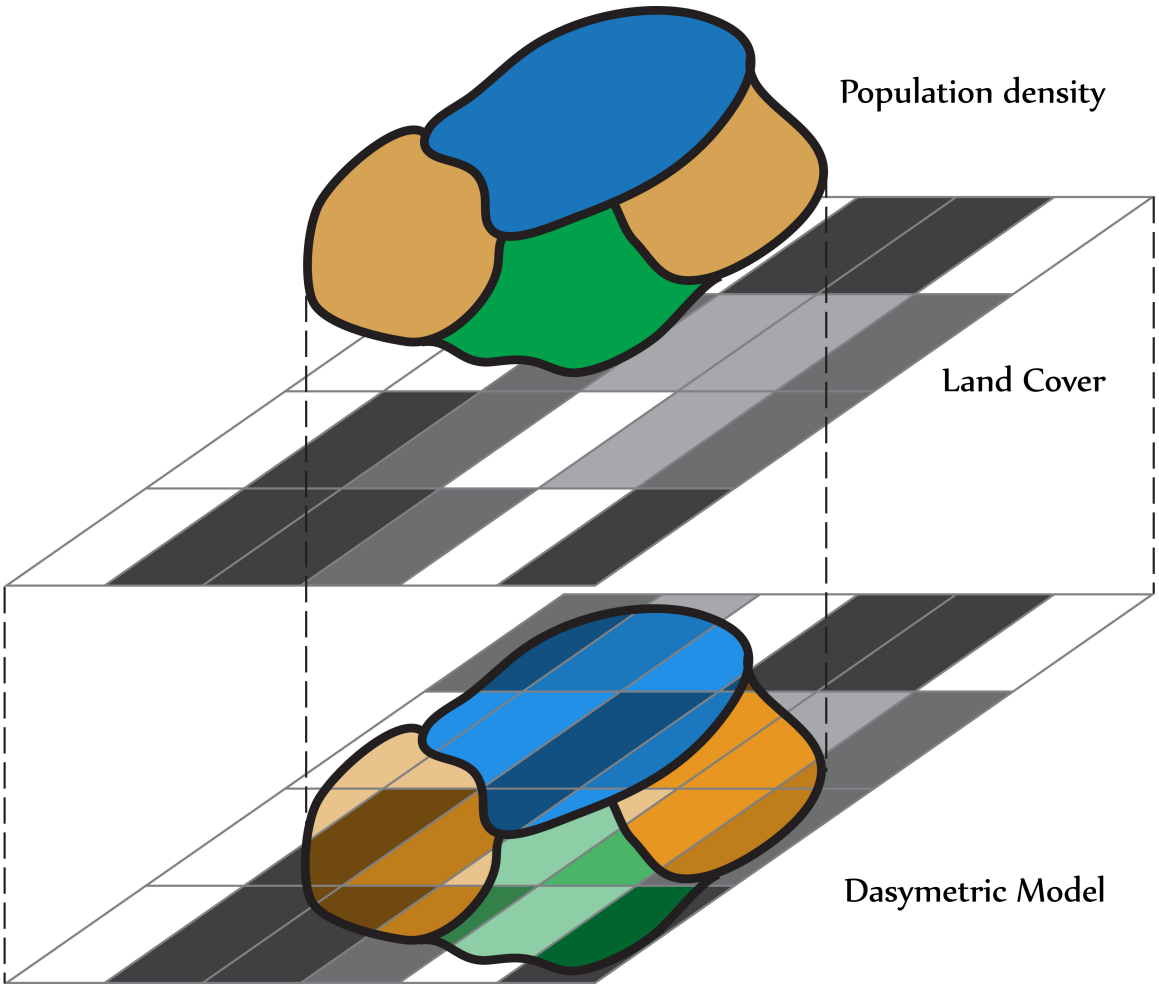


Figure 2

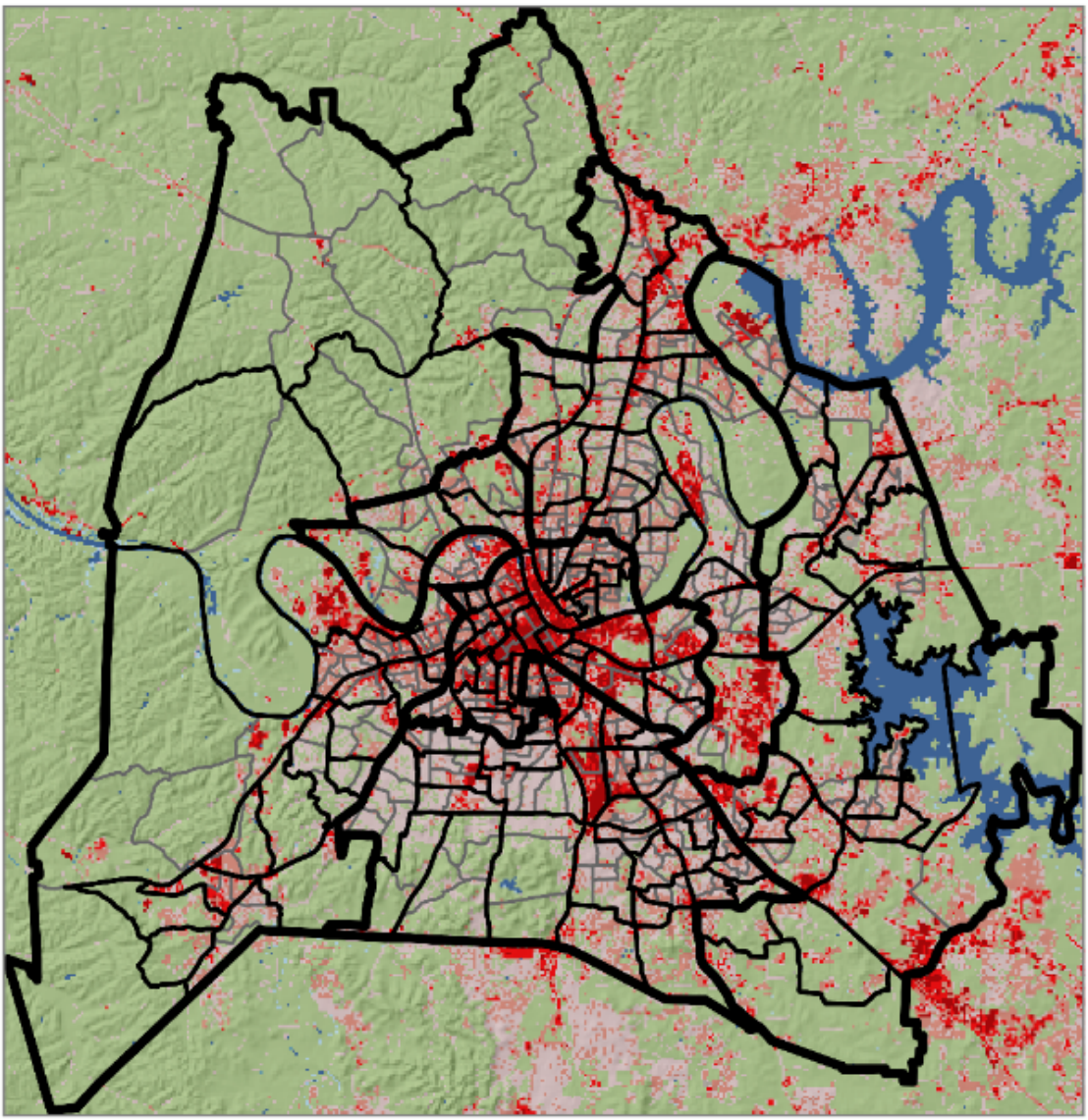


Figure 3

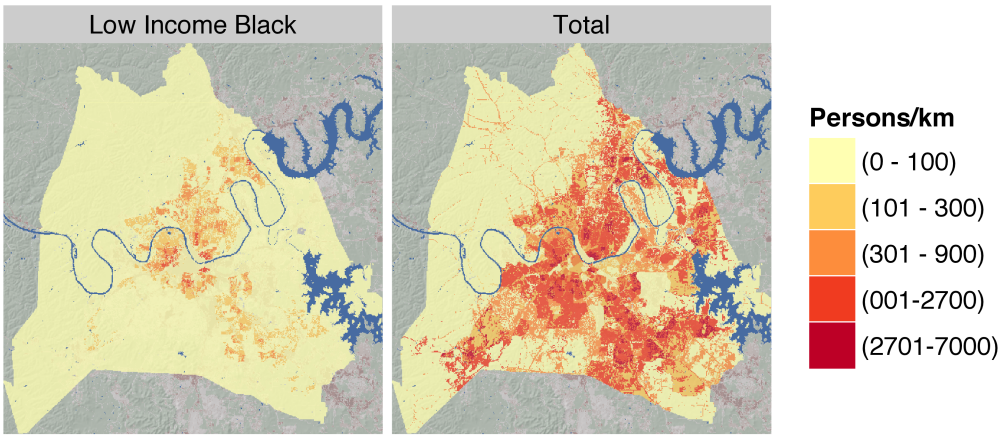

Low Income Black
Total
Persons/km
(0 - 100)
(101 - 300)
(301 - 900)
(001-2700)
(2701-7000)

Figure 4

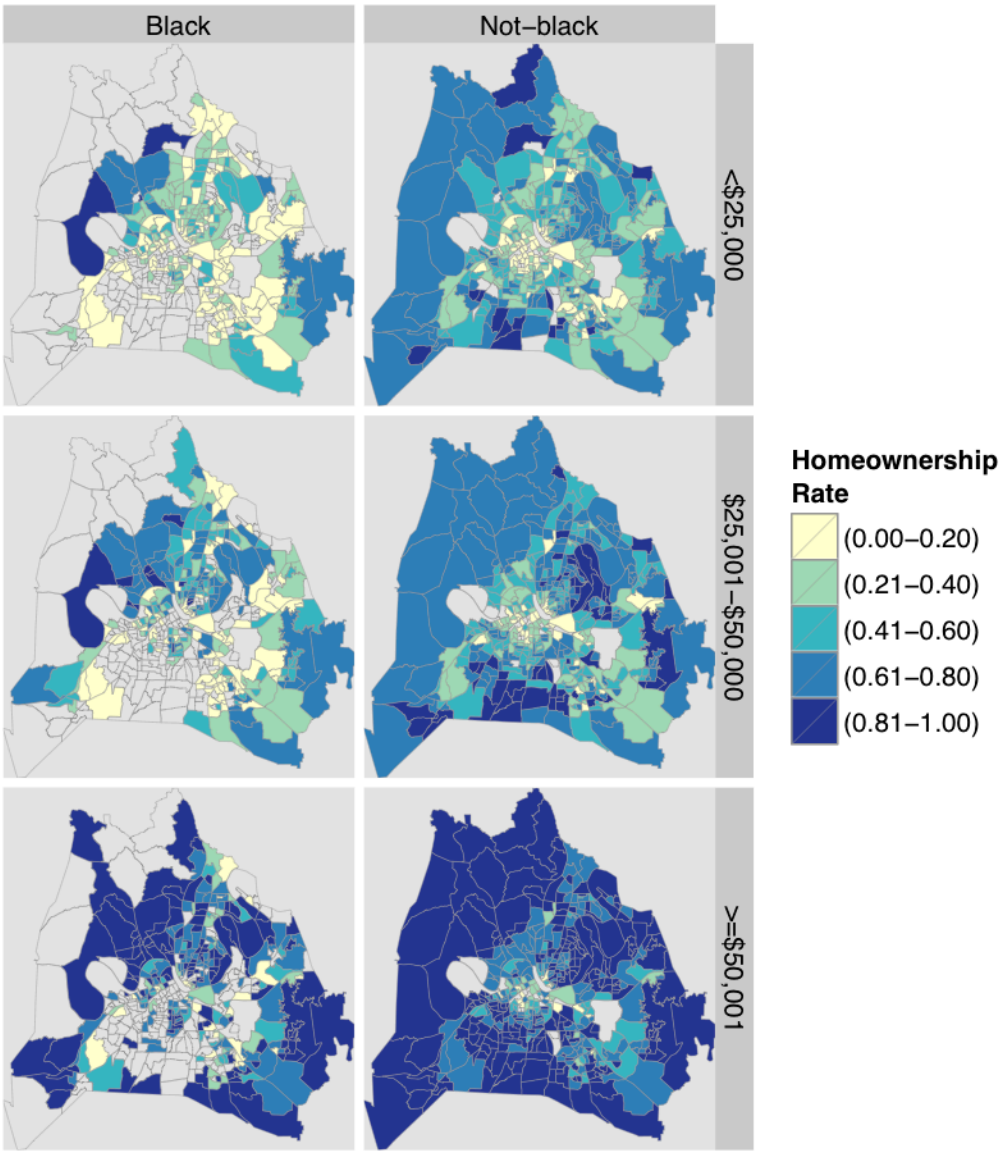
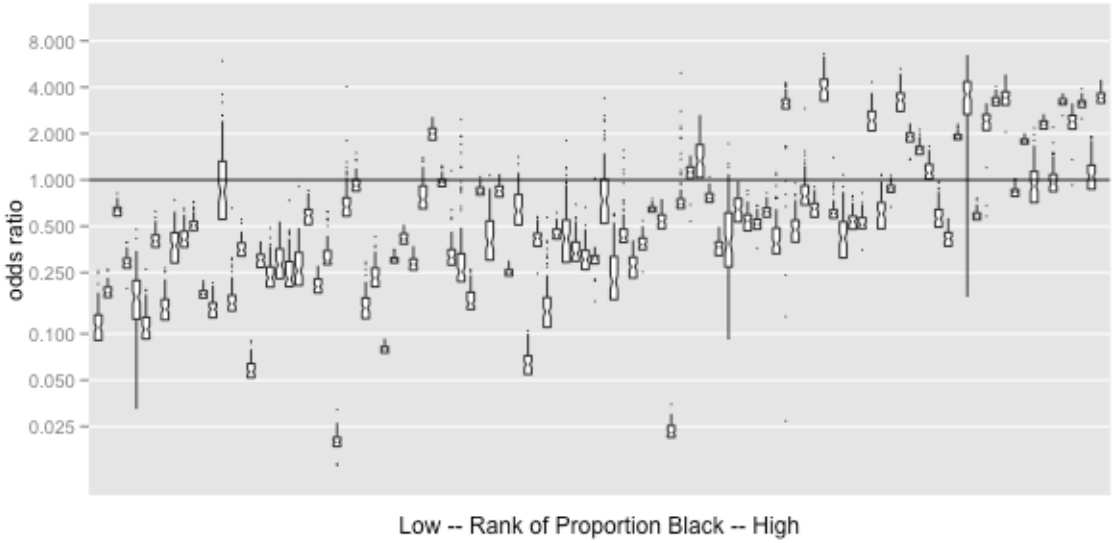

Black
Not−black
<$25,000
$25,001−$50,000
>=$50,001
Homeownership Rate
(0.00−0.20)
(0.21−0.40)
(0.41−0.60)
(0.61−0.80)
(0.81−1.00)

Figure 5


8.000
4.000
2.000
1.000
0.500
0.250
0.100
0.050
0.025
odds ratio
Low -- Rank of Proportion Black -- High