Wisconsin State Cartographer's Office

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Small-Area Population Estimation Based on Dasymetric Mapping Techniques

SCO Technical Paper

Version History

Version	Date	Notes
1.0	Dec. 2023	Initial document created.

ACKNOWLEDGEMENTS

This project was funded by the Wisconsin Coastal Management Program and the National Oceanic and Atmospheric Administration, Office for Coastal Management under the Coastal Zone Management Act, Grant # NA22NOS4190085.

The dasymetric methods described in this document were developed and implemented by Christina Dennis and Howard Veregin of the Wisconsin State Cartographer's Office at the University of Wisconsin-Madison, in partial fulfilment of the terms of the above grant.

Some of the background text in this document is from an unpublished 2016 report written by Howard Veregin, Tim Kennedy (then at the University of Wisconsin Stevens Point) and Mitch Johnson (then a student at the Wisconsin State Cartographer's Office).

INTRODUCTION

This report describes the methods used to estimate small-area populations using dasymetric mapping techniques. These population estimates are part of a study designed to improve flood risk exposure and vulnerability assessment by providing population estimates at a finer spatial scale than is possible using Census data. The dasymetric mapping techniques described here allow for the disaggregation of Census population tabulations based on publicly available parcel data.

Risk is usually conceptualized as the product of three factors: hazard, exposure and vulnerability. For flood events, hazard refers to the extent and depth of flood waters, i.e., the physical manifestation of the flood over space. Exposure and vulnerability, on the other hand, relate to the impacts of the flood on the environment. These impacts may be tangible or intangible, and may include almost anything that is considered to be of value. For example, there may be impacts on human and social systems (displacement of residents, disruption of school and work patterns, etc.), the economy (destruction of homes and property, supply chain interruptions, unemployment, etc.), infrastructure (damage to roads and culverts, impacts on power generation, etc.), the environment (effects on critical habitats and biodiversity, etc.), and government functions (loss of tax revenue, costs of infrastructure repairs, lag in emergency response times, etc.).

Exposure and vulnerability (EV) data are part of an integrated approach essential for effective flood planning, response and mitigation. Mapping of EV in the context of floods is necessary to provide reliable estimates of the social, economic and environmental resources that might be impacted. Since EV is inherently complex and multi-dimensional, a wide array of data must be collected in order to encompass all dimensions of the problem. Important dimensions include buildings and infrastructure, human populations, social and demographic variables, economic factors, the locations of care facilities, and possibly others depending on the local context (Rufat et al., 2015). Collecting and compiling this data into a usable form can be prohibitive, especially for small local governments with limited staff and budgetary resources. This leaves local decision-makers without a reliable assessment of flood EV to guide local policy and decision-making.

Equally important is the question of spatial scale. While flood hazards (i.e., extent and depth of flooding) are usually modeled using high-resolution grid cells in a GIS environment, EV data is rarely treated this way because it is usually tabulated only for large areal units. This is especially true in rural areas, where even the smallest areal unit used for Census Bureau population data – the block – can be as large as a square mile. Such areal units are too large to be used for detailed local decision-making and planning.

As shown by Rufat et al. (2015), local context and situational variability are important mediators of EV. Drivers of EV may have different effects in different contexts and may contribute to high levels of EV in one context while detracting from it in another. The importance of local context can only be accounted for through the use of localized data. Stated another way, local decision-making needs to be informed by EV data at a detailed spatial scale, not by global values and regional averages.

This is an area where geographers and cartographers can contribute. Long-standing geographic and cartographic methods such as dasymetric mapping can be used to disaggregate data from larger areal units to a more appropriate spatial scale. This process results in more detailed maps that more closely match the scale of the phenomenon being analyzed. Dasymetric mapping conventionally uses an underlying controlling (or limiting) variable such as land use to reapportion data tabulated for large areal units.

Such methods are not unknown in the risk mapping community. For example, researchers have used dasymetric mapping to estimate populations at the parcel level to provide more accurate assessments of populations at risk from rising sea levels (Mitsova et al., 2012). Similar methods have been used to develop detailed population and economic data for assessing flood risk (e.g., Amadio et al., 2019). However, these studies are the exception rather than the rule. Decision-making tools and data currently in use in Wisconsin, for example, whether at the local level or statewide (e.g., the Department of Health Service's RAFT tool) generally do not employ disaggregated spatial data within the mix of layers available for EV assessment. And while models like Hazus provide dasymetric data for flood loss modeling, the dasymetric implementation simply involves clipping census tracts to eliminate areas where people do not live, such as wetlands and water bodies; down-sampling and disaggregation of census data is not considered.

Dasymetric mapping (and similar techniques) can be used to produce detailed geospatial data for decision-makers at the local level as they plan for or respond to floods and other hazards. For example, by coupling EV data with flood hazard scenarios derived from H&H (Hydrologic & Hydraulic) models or Hazus, it is possible to quantify the impacts of a flood on people, structures, critical infrastructure, and other dimensions of EV. Moreover, it is possible to map these impacts on a detailed spatial scale to determine which areas are at the highest risk. EV data could also be used in more complex spatial models, such as assessing how access to hospitals or other critical facilities might be impacted regionally through flood impacts on the transportation network.

One practical issue that must be considered is that the methods and tools used to produce EV data must be reproducible within the constraints of time, expertise and budget faced by the agencies responsible for database development and maintenance. In the context of local governments in rural parts of Wisconsin, this implies the adaptation of existing software and workflows rather than the addition of new software extensions and complex new processes. Source data must be readily available, and processing steps clearly explained in documentation so that data can be updated and analyzed easily by staff with GIS training. If existing tools and workflows can successfully be adapted to incorporate spatially detailed EV data, the net result will be decisions that are better aligned to local context and variability, and more accurately reflect the needs of local citizens and communities.

SCOPE

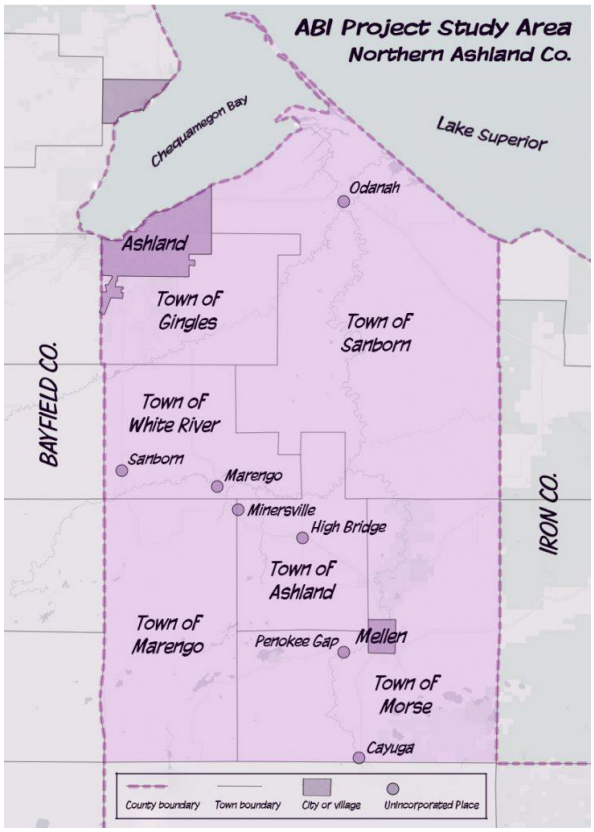
This report details the methods used to develop small-area population estimates using dasymetric mapping techniques to down-sample and disaggregate data tabulated for larger census tabulation units. The goal is to derive data at a detailed spatial scale appropriate for flood risk assessment.

The Census Bureau has released its 2020 “redistricting” data including demographic variables for census blocks. This data is available through the Geodata@Wisconsin geoportal (Esri, 2020). In this study, disaggregation of this data will be achieved using data from the Wisconsin statewide parcel database, which is available online (https://www.sco.wisc.edu/parcels/data/).

We explore several dasymetric approaches, including multiple regression analysis and non-linear optimization tools, in which the population of the block serves as the dependent variable, and the number of parcels belonging to different property classes within the block serve as the explanatory variables. The resulting model coefficients represent the contribution of each class to the population. Our final model is a spatially weighted one in which the model is iteratively fit to the closest blocks for each target block.

The study area is northern Ashland county and includes the cities of Ashland and Mellen, and the Towns of Gingles, White River, Marengo, Ashland and Morse. This area has a history of recent flooding, with federal disaster declarations in 2012, 2013, 2016 and 2018. Along the region’s Lake Superior coast, where much of the population is concentrated, there are also threats of coastal flooding associated with storm surges, lake level fluctuations, and events like seiches and meteotsunamis. As noted in the 2018 Ashland County Hazard Mitigation Plan, “there is a very high probability of flooding in the future and a very high probability of damage and losses due to flooding” with potential vulnerabilities that include residential structures, businesses and “flooded public facilities and schools, many of which are the community’s shelters needed when individual housing is uninhabitable.” The study area is predominantly rural, like most of the state. Its socio-economic rankings are generally below state averages reflecting broader trends of rural depopulation and economic decline.

The permanent impacts of this study go beyond northern Ashland county and result from the transferability of our methods to other communities. Our goal is not just to produce a single set of population estimates, but to develop a methodology that can be adapted to the specific needs of any locale. Our focus is on Wisconsin coastal communities facing risks from riverine and coastal flooding. However, the methods and tools are also applicable in other risk contexts and in other regions of the state.



Map of study area. The Bad River Reservation in the north-east part of the study area, was excluded from the dasymetric analysis, due to the complexities of land tenure in this area.

DASYMETRIC MAPPING

The introduction of dasymetric modeling to English-speaking scholars is usually credited to Wright (1936), who created a population map of Cape Cod, Massachusetts, by reapportioning town population based on tract area and population density through an examination of topographic maps and other evidence. Dasymetric modeling attempts to reapportion population statistics from a set of source polygons to a different set of target polygons based on a "controlling" or "limiting" variable thought to influence the spatial distribution of population. In effect, the controlling variable is a proxy for the unknown population density distribution of the target polygon set. The target polygons are often smaller than the source polygons, but this is not a requirement.

Dasymetric modeling is often applied when the polygons for which population statistics are available – such as census tracts – do not match the polygons required for a specific mapping or analysis application (Tapp, 2010). Our model operates by reapportioning Census Bureau block-level population statistics to individual parcels using the property classes present within these parcels as a controlling variable. Property classes define the various uses of land within parcels, such as residential, commercial and industrial.

Population estimation at the parcel level has not been a traditional focus of dasymetric modeling studies. However, small-area population estimates and projections have a variety of uses and are increasingly in demand by government agencies, researchers, and planners.

- The Southwest Florida Water Management District uses small-area population projections to support water supply planning and water use permitting (Doty, 2013).
- The San Diego Association of Governments uses small-area population estimates for a variety of purposes, including infrastructure planning, public safety, public health, and modeling accessibility (Jarosz, 2008).
- Mitsova, Esnard, and Li (2012) use parcel-level populations for Miami-Dade County, Florida, to provide assessments of populations at risk from rising sea levels associated with climate change.
- Small-area population estimates provide detailed information on where people are likely to be located when emergencies occur (Sleeter & Wood, 2006).

In general, small-area estimates are valuable because of their geographic detail and precision, which allows for greater flexibility when performing administrative, research, and planning functions (Jarosz, 2008; Tapp, 2010).

MODEL FORMULATION

Our goal is to disaggregate census polygon (block) population by assigning the population to parcels that fall within the polygon. To simplify matters and avoid problems with slivers and gaps, we use the parcel centroid, such that a given parcel can fall into only one block. The calculation is done over a set of contiguous blocks covering the study area. While we use blocks in this study, the methods would also work with block groups, tracts, or any other areal unit, except that as the size of the unit grows, disaggregation becomes less effective.

An important caveat to our method is that we want to disaggregate to *selected* parcels, not every parcel in the study area. Parcels classified as residential, for example, are likely to contain inhabitants, while those classified as agricultural or industrial probably do not. We use the term "target parcels" for the parcels we want to disaggregate to, and the term "source polygons" for the polygons we are disaggregating from.

Initially, we will assume that the target parcels we want are those with a residential property class – possibly mixed with other property classes such as agricultural – and an improved value that is greater than zero. The latter constraint means that the target parcels will contain an improved structure (a building) to differentiate them from unimproved parcels that may be classified as residential but do not contain inhabitants.

Household Size

The simplest way to disaggregate population data involves manipulation of census variables tabulated for the source polygons (blocks). Mean household size for a given polygon *i* is simply the population of the polygon divided by the number of occupied housing units in the polygon.

Thus $\hat{p}_j$ (the estimated population of target parcel *j* whose centroid falls within source polygon *i*) can be computed as:

$$\hat{p}_j = \quad p_i \,/\, h_i \qquad \text{for that value of } i \text{ where } I(j)_i = 1 \qquad (1)$$

where:

$p_i =$ known population of source polygon i

$h_i =$ number of occupied housing units in polygon i

$I(j)_i =$ indicator function with a value of 1 if target parcel j is in source polygon i and 0 otherwise

The problem with this approach is that the number of occupied housing units for polygon *i* is not always equal to the number of target parcels within *i*, i.e.,

$$h_i - \sum_{j=1}^{J} I(j)_i \stackrel{?}{=} 0 \qquad (2)$$

where:

$J =$ number of target parcels

How prevalent is this problem and what is its magnitude? In our study area in northern Ashland county, the mean of the values computed for equation (2) is 0.38. In other words, the number of occupied housing units in each block is on average slightly *greater* than the number of target parcels. Also, there is a substantial amount of variation in the relationship between occupied housing units and target parcels, given the standard deviation of 6.36 against a value of 5.61 for the mean number of occupied housing units.

Why do these discrepancies occur? There are several reasons.

1. A difference in the reference date of the parcel data and census data. Residential units might have been added or removed between the two dates.
2. Errors in the data.
3. The presence of apartments. In parcel data, an apartment will appear as a single target parcel while in census data there may be multiple housing units associated with this parcel. From the perspective of the number of target parcels, mean household size will be too large.
4. The presence of group quarters, e.g., college dorms. This case is similar to apartments, but the census reports zero housing units associated with these group quarters.
5. The presence of population in non-residential parcels. Some parcels that are not coded as residential may in fact have inhabitants. An example is mobile home parks, which may be coded as commercial. If we allow population to fall only into residential parcels in the model, we will be left with a large leftover population with no parcel to assign it to.
6. The presence of population in tax-exempt parcels. An example is a parcel owned by a municipal housing authority. Such parcels contain housing units, but in parcel data they are coded as exempt rather than residential. To complicate matters, not all exempt

properties have inhabitants, since exempt properties also include schools, churches and municipally owned properties.

7. The error intentionally introduced by the Census Bureau for disclosure avoidance. The number of households within a block is always correct, but population is purposefully manipulated. This deliberate introduction of error occurs at all levels of census geography up to the state level.
8. The distinction between occupied and unoccupied housing. The census differentiates between these, but occupancy status cannot be determined from parcel data. Many homes could be lake cottages or second homes, but our process would still treat them as target parcels.

If we use h_i (the number of occupied housing units) in conjunction with our own target parcel centroids, these errors will come into play. How do we avoid this problem? An obvious idea is to replace h_i with the number of target parcels in *i*, i.e.,

$$\hat{p}_j = p_i \,/\, \sum_{j=1}^{J} I(j)_i \qquad \text{for that value of } i \text{ where } I(j)_i = 1 \qquad (3)$$

The problem here is that equation (3) does not resolve the problem of why there should be a difference between h_i and the number of target parcels. This discrepancy is due in part to issues #4 and #6 above, where non-residential property classes contain population. Different property classes are unlikely to have the same weight. For example, the population residing on a single parcel containing dozens of mobile homes will be much greater than the number of people residing in a single-family home.

Property Class

A more general population allocation model would account for these differences in property classes. First, let us define a set of weights for a given property class *k* as follows:

$w_k =$ weight assigned to property class k

K = number of different classes of property

The value w_k is an estimate of the mean number of people per target parcel for class *k*. We will discuss later how to estimate these weights. Given a set of *K* values of w_k we can estimate target parcel population $\hat{p}_j$ simply as w_k when c_j (the property class of target parcel *j*) is equal to *k*.

However, since the w_k values are averages for the whole study area, individual target parcels may end up over- or under-estimated. This in turn implies that the sum of the populations of all target parcels within a given source polygon may not be equal to the known population of the polygon. Thus we may add or lose population in the study area, which violates the pycnophylactic constraint of density preservation. To preserve population (and density) we need to normalize all estimates. For a given polygon *i*,

$$\hat{p}_j = p_i \frac{w_{c_j}}{\sum_{j=1}^{J} \sum_{k=1}^{K} w_k I(j)_{ik}} \qquad \text{for that value of } i \text{ where } \sum_{k=1}^{K} I(j)_{ik} = 1 \qquad (4)$$

where:

c_j = property class of parcel j

K = number of unique property classes

J = number of target parcels in the study area

w_{c_j} = weight of property class c_j

w_k = weight of property class k

$I(j)_{ik}$ = indicator function with a value of 1 if target parcel j is in source polygon i and has a property class of k, and 0 otherwise

This more general formulation of the dasymetric model allows for the possibility that parcels other than residential ones can contribute population. Equation (4) lets the data "speak for itself" by allowing any property class to contribute.

Model Stratification

An even more flexible model allows the weights w_k to vary, not just with property class, but over the study area. Some regions might have different sets of weights than others. For example, it seems likely that residential parcels in a city will have different densities on average than residential parcels in a rural area. We experimented with stratified models in two ways: stratify by the mixture of property classes with each block (an ecosystem approach) or stratify by location (often called a spatially weighted approach).

In ecosystem stratification we first identify a finite number of property class mixtures that typify the study area, using some type of clustering method to group similar polygons together as representative members of the same ecosystem. For example, we might identify polygons that are dominated by improved residential parcels, polygons that contain residential parcels adjacent to agricultural land, etc. The typology can be based on known or hypothesized relationships between population density and different land use patterns, as revealed in the mixture of property classes present. To some degree, this will be based on empirically observed counts of polygons with various property class mixtures, which may vary from case to case in different study areas.

We modify equation (4) slightly to add this additional flexibility:

$$\hat{p}_j = p_i \frac{w_{c_j r_j}}{\sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{s=1}^{S} w_{ks} I(j)_{iks}} \quad \text{for that value of } i \text{ where } \sum_{k=1}^{K} \sum_{s=1}^{S} I(j)_{iks} = 1 \quad (5)$$

where:

$w_{c_j r_j}$ = weight of property class c_j within stratum r_j

c_j = property class of parcel j

r_j = stratum that parcel j belongs to

S = number of strata

w_{ks} = weight of property class k within stratum s

$I(j)_{iks}$ = indicator function with a value of 1 if target parcel j is in source polygon i,

and has a property class of k, and belongs to stratum s, and 0 otherwise

In addition to the ecological approach, we also define strata spatially, in which case we assign each parcel to one of a finite number of geographic regions. This approach explicitly accounts for spatial nonstationarity. Some research on nonstationarity in dasymetric modeling has already been carried out (Bielecka, 2005; Langford, 2006; Lin, Cromley, & Zhang, 2011; Schroeder & Van Riper, 2013). This approach is also known as a spatially-weighted approach.

ESTIMATION OF WEIGHTS

The derivation of appropriate weights to support dasymetric calculations is perhaps the most complex part of the process.

Limiting/Controlling Variables

Some studies of dasymetric mapping assign weights by distinguishing between uninhabited and inhabited areas. As noted by Reese-Cassal (2007), a common strategy to implement this binary model is to use land cover data derived from remote sensing imagery, such as the National Land Cover Database (Homer, Fry, & Barnes, 2012). In dasymetric mapping, land cover is often used as the controlling variable that drives the spatial distribution of population. In the binary model, various land cover classes are collapsed into two categories representing inhabited and uninhabited land, and weights of 1 and 0 are then assigned to their categories. Examples of this approach include Langford and Unwin (1994), Holt, Lo, and Hodler (2004), Langford (2007), Langford (2013), and Buttenfield, Ruther, and Leyk (2015).

The binary model has the advantage of simplicity, but – as noted in the Model Formulation section above and by other researchers (e.g., Langford, 2006) – it is unable to account for variations in density associated with more complex mixtures of land cover classes.

To accommodate additional land cover classes, target parcel (or more generally target polygons, since some studies are not parcel-based) weights must be allowed to take values other than 0 or 1. For example, high-density residential areas might be expected to contribute more population per square mile than low-density areas; to adequately model this fact three or more classes are needed, each with its own weight. In such models, weights vary among classes but are uniform within any one class, and as such can be viewed as average population densities for each class. This multi-class approach is quite popular (Eicher & Brewer, 2001; Giordano & Cheever, 2010; Jia, Qiu, & Gaughan, 2014; Mennis, 2003; Tapp, 2010). Some researchers have noted that the expected performance gains of the multi-class model do not always emerge empirically; however, research on this question is far from conclusive (Langford, 2006).

When more than two classes are involved, a method must be found to specify a weight for each class. Some researchers have taken a subjective approach to this problem. Wright (1936), for example, used “educated guesswork” – which raises concerns about accuracy and repeatability. Most dasymetric models are empirically based, with weights dependent on the controlling variable. In the case of land cover, for example, it is common practice to identify multiple classes of developed land – such as high, medium, and low intensity – and then assign each class a specific weight.

For Eicher and Brewer (2001) and Mennis (2003) estimation of class weights involves the identification of source polygons that are categorically pure, i.e., that contain only one class. The observed population densities of these pure polygons are then used to compute the weight for the class. A problem with this approach is that pure polygons tend to be rare in many datasets, which leads to questions about how representative such polygons can be of the whole dataset. A related problem is that to identify a sufficiently large sample, it is often necessary to relax the rules and use polygons that are not perfectly pure (Giordano & Cheever, 2010; Jia et al., 2014; Mennis & Hultgren, 2006). An obvious question is how far the assumption of purity can be relaxed before accuracy suffers. Moreover, the likelihood of finding pure polygons diminishes as the number of classes increases, making this approach even more problematic for more complex dasymetric models.

Despite the popularity of land cover as a controlling variable, there is an inherent fallacy associated with using land cover in this context. The land cover classes derived from remote sensing imagery correspond to mixtures of physical features – streets, rooftops, tree canopies, etc. – that give rise to characteristic spectral response profiles identifiable on imagery. However, these land cover classes are an inadequate surrogate for the various land use categories representing human habitation of the land. For example, the land cover class "developed" typically includes not just the apartments, condos, and single-family homes where people reside, but also warehouses, highways, parking lots, and other features not normally used as residences. Dasymetric population maps based on land cover often depict such areas erroneously as having significant population concentrations (Hackett, Veregin, & Cox, 2015; Zandbergen & Ignizio, 2010).

Parcels

Only a few researchers have focused on dasymetric modeling at the parcel level. Sleeter and Wood (2006) used two parcel attributes – land use and building type – to partition parcels into four density classes: high, medium, low, and uninhabited. Population densities for these classes were then estimated using a sampling approach. Other researchers have derived weights based on the number of address points or the number of residential housing units within each parcel (Mitsova et al., 2012; Tapp, 2010).

Some researchers view parcels as a data source rather than a destination for population estimates. For example, Jia, Qiu, and Gaughan (2014) and Jia and Gaughan (2016) used parcel data to access residential property tax categories, which were then used in a model to estimate populations for a gridded dataset. In this approach, parcels are important because they improve the accuracy of the final gridded dataset, rather than because of any intrinsic interest in population estimates for the parcels themselves.

The Southwest Florida Water Management District uses two main variables to estimate parcel populations: the total number of existing residential units (households) within each parcel, obtained from a parcel dataset, and average population per household (household size), derived at the tract level from Census Bureau statistics. The estimated population of a parcel is then computed as the number households in the parcel multiplied by the average household size (Doty, 2013).

Empirical Derivation of Weights

Many researchers agree that sampling and statistical analysis is necessary to estimate class weights and that regression analysis provides an appropriate framework. In this approach, weights are computed empirically based on observed relationships between source polygon populations and their class compositions. Flowerdew and Green (1989) pioneered the use of Poisson regression in this context under the assumption that population follows a Poisson distribution. Other researchers have rejected Poisson regression in favor of Ordinary Least Squares (OLS) regression, in part because of OLS's computational simplicity (Reibel & Agrawal, 2007). The main problem with OLS regression in this context is that it does not guarantee that coefficients will be non-negative, which may give rise to negative population densities (Langford, 2006; Yuan, Smith, & Limp, 1997). Some researchers have suggested modifications to OLS regression to ensure that population estimates are always positive (Goodchild, Anselin, & Deichmann, 1993).

There are some useful parallels here to hedonic modeling as used in econometrics. The premise of hedonic modeling is that the price of a marketed good is related to its characteristics, and that it is possible to attach a value to individual characteristics by parameterizing the price people are willing to pay for them (Bell & Irwin, 2002; Rosen, 1974). Statistically this amounts to regressing the price of a good, such as a house, on a set of characteristics like house size, the number of bedrooms, and the presence of an attached garage. The resulting regression coefficients represent how much consumers are willing to pay for each of these characteristics. The total price of the house is a linear combination (sum) of the weight (in dollars) of each characteristic.

To apply this method to dasymetric modeling in our study, we substitute source polygon (census block) population for house price and the count (number of parcels) of each property class within the source polygons in place of housing characteristics. The resulting regression coefficients represent the total population contributed per parcel for each class. The model can be expressed as follows.

$$\dot{p}_i = \sum_{k=1}^{K} b_k \, q_{ik} \qquad (6)$$

where:

$\dot{p}_i$ = estimate of population for source polygon i

b_k = weight (regression coefficient) for class k

q_{ik} = count of property class k parcels within source polygon i

K = number of unique property classes

The coefficients b_k represent population density values for each class, and as such they can be used in equation (4) – as well as equation (5) with some modification. Note that in this context regression analysis is not being carried out to test hypotheses, but rather as a practical solution to a problem of estimation (Reibel & Agrawal, 2007).

Equation (6) can be implemented using OLS regression; however, one problem with OLS is that it does not constrain coefficients to be positive, which can lead to negative population estimates. Our solution is to use Generalized Reduced Gradient (GRG) optimization in place of OLS. GRG is an advanced analytics tool that can be used to solve a variety of nonlinear problems and is available in desktop software including the Microsoft Excel Solver tool.[1] In Excel, GRG is easy to set up and runs quickly, and results can easily be incorporated into other equations and calculations.

For dasymetric modeling, the GRG model is formulated with the goal of minimizing the sum of squared deviations between actual and estimated populations of source polygons. The model modifies class weights until a solution is attained. The formal specification is as follows:

$$\text{minimize} \quad \sum_{i=1}^{I}(p_i - \dot{p}_i)^2 \tag{7}$$

$$\text{such that} \quad b_k \geq 0, \quad \forall\, k$$

where p_i = observed population of source polygon i

$\dot{p}_i$ = estimate of population for source polygon i

b_k = weight for property class k

I = number of source polygons

Note that this formulation imposes a constraint that the weights for all property classes, b_k, are non-negative.

One limitation of GRG and other nonlinear optimization tools is the existence of multiple local feasible solutions where all constraints are satisfied. Generally, there is no way to determine which local solution is globally optimal. Because of this problem, it is a good idea to run a GRG model using starting values based on *a priori* knowledge to increase the chances of finding the optimal solution. We use standard OLS estimates as the starting values.

Property Class

The controlling variable in our case study is a property class attribute used for tax assessment at the parcel level. This attribute identifies seven primary property classes and a variety of secondary or auxiliary classes and is closely associated with human habitation. We anticipate that the residential class will contribute the most to population loadings for parcels, although additional classes, such as the other class, also have population associated with them. Property class and auxiliary property class definitions can be found here: https://www.sco.wisc.edu/parcels/data/assets/V9/V9_Wisconsin_Statewide_Parcels_Schema_Documentation.pdf

MODEL IMPLEMENTATION

We employed a variety of tools and software packages to explore dasymetric models for this study. Here we report only on our final and best result, a spatially weighted GRG solution

[1] Our use of Microsoft Excel should not be construed as an endorsement of this particular software product. Other implementations of GRG also exist.

implemented in Microsoft Excel using standard tools including the Solver, Pivot Tables, the vlookup function and Macros.

The major steps in the implementation are as follows.

1. Simplify the property class classification to reduce the number of possible combinations of property class and auxiliary property class for parcels. Since a parcel may contain mixtures of these classes, the number of possible combinations can grow quite large, with some combinations occurring only rarely. The simplification process is necessarily somewhat subjective. In our case, we created six final classes: RES (parcels containing a residential property class, possibly in conjunction with other classes, with an improved value greater than zero); RES0 (parcels containing a residential property class, possibly in conjunction with other classes, but with improved value equal to zero); COMM (parcels classified as commercial with no residential component); AUXX (parcels classified as Auxiliary Class X, usually denoting public ownership or other tax-exempt status), AUXW (parcels classified as Auxiliary Class W, associated with managed forest land), and OTHER. Each parcel is assigned to one and only one of these six classes.
2. Use GIS tools to compute the number of parcel centroids of each of the six classes within each census block, and to populate a cross-walk between parcel IDs and block IDs.
3. Create a table in Excel where each row represents a block. Important attributes to include in this table are: a) block ID; b) block population from census; c) count of each class computed in #2 above, with one column for each class; d) six empty fields that will hold the GRG-derived coefficients, one for each class; e) the latitude and longitude (or easting and northing) of each block centroid; f) column to hold the distance between the block centroid and the "target block" that will be estimated; g) column to hold the initial population estimate for the block, computed from the weights derived from the GRG analysis multiplied by the parcel count for each class, i.e., equation (6); h) column to hold the final pop estimate; i) a pre-estimate squared error column, computed as the square of the difference between the initial population estimate and the census population; and j) a post-estimate squared error column, computed as the square of the difference between the final population estimate and the census population.
4. The table also needs to have a cell with the objective function for the Solver, in this case the sum of squared errors for the first fifty rows of the table. The first fifty rows are used because we are using the closest fifty blocks to the target block. The sum of squares is the sum of the first fifty rows of the pre-estimate squared error column (*i* in #3 above). The solver will try to minimize this value.
5. The table needs six blank cells that will hold the Solver-derived coefficient estimates.
6. Compute the distance from the first block in the table to every other block and store the result in the distance column. Use the Pythagorean Theorem for eastings/northings and great circle distance for latitude-longitude coordinates. Next, a) sort the rows of the table from smallest to largest distance; b) initialize the Solver coefficients to 1; c) run the Solver to minimize the error sum of squares for the closest fifty blocks, which will also

compute the initial population estimate for the block; and d) store the coefficients from the solver (d in #3 above) and the population estimate (h in #3 above).

7. Repeat the procedure in a loop for every block in turn. This was implemented as a macro in Excel. (Macros and scripts are two different ways of automating procedures in Excel.)
8. Create a second table in Excel where the rows are the parcels. Important attributes are a) the parcel ID, b) the block ID of the parcel, c) the simplified class of the parcel (one of six possible values), d) a column for the GRG coefficient for the parcel; e) a column for the denominator of the normalization equation as in equation (5); and f) a column for the final population estimate of the parcel.
9. Using the Excel vlookup function, assign the correct GRG coefficient to each parcel based on its simplified class and its block membership.
10. Using Pivot Tables, calculate the denominator of the normalization equation.
11. For each row of the parcel table, divide the GRG coefficient by the denominator value to compute the final population estimate for the parcel.
12. Results can be extracted to a text table, joined to the parcel feature class, and used for mapping and analysis purposes.

RESULTS

Compared to other models we experimented with, the spatially weighted Solver model yielded a higher pseudo-R^2 value, of about 0.75, compared to other models that were below 0.5. This value reflects the relatively good model fit when we allow coefficients to vary spatially over the study area. The map of population estimates is part of an online app developed for the project.

Our results show that population estimates at the parcel level can be obtained using dasymetric techniques using standard GIS and office software. The Excel model developed here could be converted to another platform or language depending on users' skills and preferences. The approach requires parcel data with an attribute describing the classes of property and, ideally, improved value. This study used block-level census data, but other levels of census data could also be used.

While the high pseudo-R^2 value of the spatially weighted Solver model indicates a good fit to the data, we cannot easily establish a quantitative value for the accuracy of the parcel estimates themselves. This cannot be assessed without reference to independent population data at the parcel level.

REFERENCES

Amadio, M., Mysiak, J., & Marzi, S. (2019). Mapping socioeconomic exposure for flood risk assessment in Italy. Risk Analysis, 39(4), 829-845.

Bell, K. P., & Irwin, E. G. (2002). Spatially explicit micro-level modelling of land use change at the rural–urban interface. Agricultural Economics, 27(3), 217-232. doi:10.1016/S0169-5150(02)00079-8

Bielecka, E. (2005). A dasymetric population density map of Poland. Paper presented at the Proceedings of the 22nd International Cartographic Conference.

Buttenfield, B. P., Ruther, M., & Leyk, S. (2015). Exploring the impact of dasymetric refinement on spatiotemporal small area estimates. Cartography and Geographic Information Science, 42(5), 449-459.

Doty, R. L. (2013). The Small Area Population Projection Methodology Used by the Southwest Florida Water Management District. GIS Associates, Inc. Gainesville, Florida.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. Cartography and Geographic Information Science, 28(2), 125-138.

Environmental Systems Research Institute, Inc. (Esri). (2020). Census Blocks with Redistricting Data, Wisconsin 2020. https://geodata.wisc.edu/catalog/367B52FB-6AEC-47F9-978D-29D1876AF057

Flowerdew, R., & Green, M. (1989). Statistical methods for inference between incompatible zonal systems. Accuracy of spatial databases, 239-247.

Foody, G. M. (2004). Thematic map comparison. Photogrammetric Engineering & Remote Sensing, 70(5), 627-633.

Geist, H. J., McConnell, W., Lambin, E. F., Moran, E., Alves, D., & Rudel, T. (2006). Causes and trajectories of land-use/cover change Land-Use and Land-Cover Change (pp. 41-70).

Giordano, A., & Cheever, L. (2010). Using dasymetric mapping to identify communities at risk from hazardous waste generation in San Antonio, Texas. Urban Geography, 31(5), 623-647.

Gobster, P. H., & Rickenbach, M. G. (2004). Private forestland parcelization and development in Wisconsin's Northwoods: perceptions of resource-oriented stakeholders. Landscape and Urban Planning, 69(2–3), 165-182. doi:http://dx.doi.org/10.1016/j.landurbplan.2003.09.005

Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. Environment and Planning A, 25(3), 383-397.

Hackett, B., Veregin, H., & Cox, T. (2015, February 19, 2015). Population Density Mapping Using the Dasymetric Method. Paper presented at the Wisconsin Land Information Association Annual Conference, Green Bay, WI.

Haines, A., Kennedy, T., & McFarlane, D. (2011). Parcelization: forest change agent in northern Wisconsin. Journal of Forestry, 109(2), 101-108.

Haines, A., & McFarlane, D. (2012). Factors Influencing Parcelization in Amenity-Rich Rural Areas. Journal of Planning Education and Research, 32(1), 81-90.

Holt, J. B., Lo, C., & Hodler, T. W. (2004). Dasymetric estimation of population density and areal interpolation of census data. Cartography and Geographic Information Science, 31(2), 103-121.

Homer, C. H., Fry, J. A., & Barnes, C. A. (2012). The national land cover database. US Geological Survey Fact Sheet, 3020(4), 1-4.

Jarosz, B. (2008). Using assessor parcel data to maintain housing unit counts for small area population estimates Applied demography in the 21st century (pp. 89-101): Springer.

Jia, P., & Gaughan, A. E. (2016). Dasymetric modeling: A hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida. Applied Geography, 66, 100-108.

Jia, P., Qiu, Y., & Gaughan, A. E. (2014). A fine-scale spatial population distribution on the High-resolution Gridded Population Surface and application in Alachua County, Florida. Applied Geography, 50, 99-107.

Kennedy, T. T., & Veregin, H. (in press 2016). Parcelization in rural agricultural and forested landscapes in Wisconsin, 1972–2007: evaluating multiple dimensions of human decision-making over time. Journal of Land Use Science. UPDATE - no longer in press.

Lambin, E. F., Geist, H. J., & Lepers, E. (2003). Dynamics of land-use and land-cover change in tropical regions. Annual Review of Environment and Resources, 28(1), 205-241. doi:10.1146/annurev.energy.28.050302.105459

Langford, M. (2006). Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. Computers, Environment and Urban Systems, 30(2), 161-180.

Langford, M. (2007). Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. Computers, Environment and Urban Systems, 31(1), 19-32.

Langford, M. (2013). An evaluation of small area population estimation techniques using open access ancillary data. Geographical Analysis, 45(3), 324-344.

Langford, M., & Unwin, D. J. (1994). Generating and mapping population density surfaces within a geographical information system. The Cartographic Journal, 31(1), 21-26. doi:10.1179/000870494787073718

Lin, J., Cromley, R., & Zhang, C. (2011). Using geographically weighted regression to solve the areal interpolation problem. Annals of GIS, 17(1), 1-14.

McGranahan, D. (1999). Natural Amenities Drive Rural Population Change. Retrieved from Washington, D.C.: http://www.ers.usda.gov/publications/aer781/

Mehmood, S. R., & Zhang, D. (2001). Forest Parcelization in the United States: A Study of Contributing Factors. Journal of Forestry, 99(4), 30-34.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping∗. The Professional Geographer, 55(1), 31-42.

Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. Cartography and Geographic Information Science, 33(3), 179-194.

Mitsova, D., Esnard, A.-M., & Li, Y. (2012). Using enhanced dasymetric mapping techniques to improve the spatial accuracy of sea level rise vulnerability assessments. Journal of Coastal Conservation, 16(3), 355-372.

Morisette, J. T., & Khorram, S. (1998). Exact binomial confidence interval for proportions. Photogrammetric Engineering and Remote Sensing, 64(4), 281-282.

Ohm, B. W. (1999). Guide to Community Planning in Wisconsin. Madison: University of Wisconsin-Extension.

Reese-Cassal, K. (2007). An evaluation of land parcel-weighted areal interpolation in small area estimates. Paper presented at the Population Association of America Conference, New York City, NY.

Reibel, M., & Agrawal, A. (2007). Areal interpolation of population counts using pre-classified land cover data. Population Research and Policy Review, 26(5-6), 619-633.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. Journal of Political Economy, 82(1), 34-55.

Rufat, S., Tate, E., Burton, C. G., & Maroof, A. S. (2015). Social vulnerability to floods: Review of case studies and implications for measurement. International journal of disaster risk reduction, 14, 470-486.

Schroeder, J. P., & Van Riper, D. C. (2013). Because Muncie's Densities Are Not Manhattan's: Using Geographical Weighting in the Expectation–Maximization Algorithm for Areal Interpolation. Geographical Analysis, 45(3), 216-237.

Sleeter, R., & Wood, N. (2006). Estimating daytime and nighttime population density for coastal communities in Oregon. Paper presented at the 44th Urban and Regional Information Systems Association Annual Conference, British Columbia.

Stone, R. S., & Tyrrell, M. L. (2012). Motivations for Family Forestland Parcelization in the Catskill/Delaware Watersheds of New York. Journal of Forestry, 110(5), 267-274.

Tapp, A. F. (2010). Areal interpolation and dasymetric mapping methods using local ancillary data sources. Cartography and Geographic Information Science, 37(3), 215-228.

Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. Journal of the American Statistical Association, 74(367), 519-530.

Wright, J. K. (1936). A method of mapping densities of population: With Cape Cod as an example. Geographical Review, 26(1), 103-110.

Yuan, Y., Smith, R. M., & Limp, W. F. (1997). Remodeling census population with spatial information from Landsat TM imagery. Computers, Environment and Urban Systems, 21(3), 245-258.

Zandbergen, P. A., & Ignizio, D. A. (2010). Comparison of dasymetric mapping techniques for small-area population estimates. Cartography and Geographic Information Science, 37(3), 199-214.