Published in final edited form as:

Geogr Anal. 2013 July 1; 45(3): 216–237. doi:10.1111/gean.12014.

Because Muncie's Densities Are Not Manhattan's: Using Geographical Weighting in the EM Algorithm for Areal Interpolation

Jonathan P. Schroeder and David C. Van Riper

Minnesota Population Center, University of Minnesota, Minneapolis, MN, USA

Abstract

Areal interpolation transforms data for a variable of interest from a set of source zones to estimate the same variable's distribution over a set of target zones. One common practice has been to guide interpolation by using ancillary control zones that are related to the variable of interest's spatial distribution. This guidance typically involves using source zone data to estimate the density of the variable of interest within each control zone. This article introduces a novel approach to density estimation, the geographically weighted expectation-maximization (GWEM) algorithm, which combines features of two previously used techniques, the expectation-maximization (EM) algorithm and geographically weighted regression. The EM algorithm provides a framework for incorporating proper constraints on data distributions, and using geographical weighting allows estimated control-zone density ratios to vary spatially. We assess the accuracy of GWEM by applying it with land-use/land-cover ancillary data to population counts from a nationwide sample of 1980 United States census tract pairs. We find that GWEM generally is more accurate in this setting than several previously studied methods. Because target-density weighting (TDW)—using 1970 tract densities to guide interpolation—outperforms GWEM in many cases, we also consider two GWEM-TDW hybrid approaches, and find them to improve estimates substantially.

Introduction

The National Historical Geographic Information System (NHGIS) is a web-based data access system that provides summary statistics and boundary data for United States (U.S.) censuses from 1790 to the present (<http://www.nhgis.org>). A follow-up project, the Integrated Spatio-Temporal Aggregate Data Series (ISTADS), is undertaking temporal integration of NHGIS data, linking together sets of comparable variables and standardizing geographic units across censuses (Noble et al. 2011). To integrate data for areas where boundaries changed between censuses, one approach the project employs is areal interpolation, which takes data describing a variable's distribution over one set of areal units (the *source zones*) and transforms them to estimate the same variable's distribution over a different set of areal units (the *target zones*). For example, U.S. census tract definitions frequently change between censuses. Many tracts are split or merged in areas of population growth or decline, respectively, and boundaries also occasionally are adjusted to conform to changes in administrative areas or road networks. Such changes complicate the linkage of tract data across time. By using areal interpolation, census tracts may be set as the target zones, and source data from other censuses may be interpolated to produce a time series for each target tract. The challenge here is to identify an interpolation method that produces satisfactory estimates.

The scope of this challenge depends on the exact census areas and years involved. When target zones are large relative to source zones, such that most source zones nest within target zones, then most source data can be allocated directly to the encompassing target zones with no interpolation error. Therefore, producing tract time series by interpolating from census blocks, the smallest census reporting areas, which generally nest well within the tracts of any census year, is desirable because interpolation is needed only for the small portion of blocks that straddle another year's tract boundaries. This approach makes producing sufficiently accurate estimates possible even when using a simple interpolation method such as areal weighting, which allocates data from source zones to target zones in proportion to the areas of intersection between them. Areal weighting's basic assumption—that densities are spatially uniform within each source zone—often may be inaccurate, but among all blocks intersecting a given tract, the portion of data that could be misallocated is typically small. Basing interpolation on block data, however, is feasible only for 1990 and later years—years for which NHGIS provides the required block boundary data. For earlier years, the smallest units with NHGIS boundary data are census tracts. Thus, to extend tract time series to earlier years, one viable alternative is to apply areal weighting using historical census tracts as source zones, which most likely will produce larger errors than block-based areal weighting.

Another alternative when constructing time series is to assume that the densities within each source zone have the same proportional distribution as the densities of intersecting target zones in the “target year.” For example, to allocate a 1980 tract's population to two intersecting 1970 tracts, where one of the 1970 tracts had twice the population density as the other *in 1970*, we could assume that the same 2:1 density ratio existed *in 1980* in the corresponding parts of the 1980 tract. This approach, target-density weighting (TDW), generally is more accurate than areal weighting (Schroeder 2007), but there are some common settings in which TDW may produce large errors (e.g., in areas of urbanization or redevelopment, where one year's density distribution may differ greatly from another year's, or in cases where a target zone is much larger than its intersection with a source zone, making the entire target zone's density a poor indicator of density within the source zone). Schroeder (2007) defines an error model that can be used to compute a probable range of error for any TDW estimate, but this calculation only provides a means to assess uncertainty. It does not improve interpolation accuracy.

In this article, we examine how a different type of ancillary data—land use/land cover (LULC) data—might be used to achieve more accurate areal interpolation of historical census tract data. Assuming a strong relationship between LULC zones and population distributions is intuitive, and numerous studies show empirically that LULC data are an effective interpolation guide (e.g., Langford, Maguire, and Unwin 1991; Fisher and Langford 1995; Holt, Lo, and Hodler 2004; Langford 2006; Reibel and Agrawal 2007; Lin, Cromley, and Zhang 2011; Cromley, Hanink, and Bentley 2012). Other types of ancillary data also have proved to be effective, such as road network data (Xie 1995; Mrozinski and Cromley 1999; Reibel and Bufalino 2005) and address points or parcel data (Tapp 2010), but few digitized nationwide historical sources of such types of high-detail ancillary data exist. However, a (nearly) nationwide digitized LULC data set that is contemporary with the 1980 census does exist, namely the U.S. Geological Survey's Geographic Information Retrieval and Analysis System (GIRAS) data, which were compiled by interpretation of aerial photographs from the mid-1970s through the early 1980s (Price et al. 2006, U.S. Geological Survey 1990).

Our specific aims in summarizing our research are to assess how best to employ GIRAS LULC data in the areal interpolation of 1980 census tract data, and to determine whether doing so produces more accurate estimates than either areal weighting or TDW, as we might

expect. Past research about areal interpolation and about the related subject of dasymetric mapping already has generated a broad range of techniques for settings like this, in which the general aim is to disaggregate zonal counts of a feature of interest (usually population), and the spatial distribution of the feature is expected to correspond to some set of “control zones” (often LULC class areas). In the case of dasymetric mapping, the end goal is to map the population distribution among control zones in order to achieve a more exact depiction than a choropleth map of the source zones. In the case of areal interpolation, the end goal is to estimate target zone counts by summing estimates for the control zones within each target zone.

Among the many established approaches to dasymetric modeling, as summarized in recent reviews by Mennis (2009) and Holt and Lu (2011), we identify the following two that we believe should be well suited to the present application setting and others like it: the expectation-maximization (EM) algorithm, and geographically weighted regression (GWR). As we subsequently describe in more detail, both approaches, in dasymetric modeling applications, estimate control zone densities by fitting statistical models that relate source zone populations to control zone areas. The EM algorithm, as specified for areal interpolation by Green (1989), iteratively refines estimates of populations and densities through a feedback loop. This framework enables it to incorporate important constraints that the more commonly used techniques of linear regression (including GWR) do not. Specifically, it yields no negative count estimates, and it maintains the pycnophylactic property (Tobler 1979), whereby the total of estimated counts for all areas within a source zone equals the original source zone count. A limitation of Green's EM specification is that it estimates the densities of different control zones to have a constant ratio among all source zones. For example, the estimated ratio between the densities of residential and industrial control zones would be the same for a source zone in Manhattan, New York, or in Muncie, Indiana.

In contrast, using GWR to estimate control zone densities allows the ratios between estimated densities to vary spatially. Instead of producing one global set of estimates, GWR produces unique local estimates by refitting a weighted regression for any number of locations, in each case assigning greater weight to nearby observations. In two previous studies using GWR for population estimation in the Atlanta metropolitan area (Lo 2008) and a Connecticut county (Lin, Cromley, and Zhang 2011), GWR generally is more accurate than unweighted regression, although not by consistently large degrees. We expect geographical weighting to provide a substantially better improvement when interpolating tract data across the entire diverse U.S.

The principal innovation we introduce and assess in this article is to combine the features of the EM and GWR techniques in a “geographically weighted expectation-maximization” (GWEM) algorithm. In subsequent sections, we provide more detail about our data and methods, and we then assess the accuracy of GWEM for tract data interpolation by using it to disaggregate population counts for a large, U.S. nationwide sample of 1980 tract pairs (aggregations of neighboring tracts). This strategy enables the measurement of errors by comparison with known 1980 tract populations. In this setting we find that GWEM generally is more accurate than areal weighting, TDW, linear least-squares regression, GWR, and the unweighted EM algorithm, and therefore should be effective in other similar settings as well. In addition, although GWEM is more accurate than TDW on average, TDW produces more accurate estimates in many cases, resulting in a lower median absolute error than GWEM. Consequently, we assess two hybrid strategies that combine GWEM and TDW estimates through a weighted average, and we find these to outperform the other estimates we include by a substantial margin.

GIRAS data characteristics and limitations

The GIRAS dataset is divided into 471 separate tiles drawn from maps originally published at scales of 1:250,000 or 1:100,000. These tiles cover most of the contiguous U.S. and Hawaii, but there are a few gaps between tiles and along the country's coasts and boundaries, and only one tile is complete for Alaska. Therefore, we omit from our analysis all 52 of the 1980 census tracts in Alaska, none of which are covered by a GIRAS tile, along with the 30 tracts in the contiguous U.S. that have more than one hectare of area lying outside any tile. Ultimately, these areas will require an alternative method of areal interpolation. Each GIRAS polygon delineates an area corresponding to one of the 37 Level II LULC classes from the Anderson et al. (1976) classification scheme. For modeling population distributions, this scheme is needlessly and somewhat problematically precise. Even if significant differences exist among the population densities of the four classes of agricultural land or five classes of tundra, these differences are impossible to estimate with confidence as many of the classes occur rarely and comprise small portions of the few tracts in which they appear.

We reclassify the data into a smaller set of classes that we expect to have important, measurable differences in population density in at least some parts of the U.S. (Table 1). This includes each of the Level II Urban or Built-Up Land classes (Residential; Commercial and Services; Industrial; Transportation, Communications and Utilities; Industrial and Commercial Complexes; Mixed Urban or Built-up Land; and Other Urban or Built-up Land), several Level I classes (Agricultural Land, Rangeland, Forest Land, Water, and Wetland), and one aggregation of three Level I classes (Barren Land, Tundra, and Perennial Snow or Ice). In addition, we keep one Level II class—Transitional Areas—separated from the rest of the Barren Land group because such areas have the potential for significant populations at the time of the 1980 census. We also reclassify all occurrences of the Commercial and Services land use into one of two classes within each 1980 tract, Small or Large Commercial, determined by whether the class covers an area less or greater than 400 hectares in each tract. This differentiation effectively distinguishes military bases and some other large, low-density zones that are classified as Commercial and Services from more typical, smaller commercial zones. The final result is a set of 16 classes, including a special class for “No Data” areas—small tract parts uncovered by GIRAS tiles or, more commonly, stretches along the U.S. borders and coasts where the NHGIS tract boundaries extend beyond GIRAS's national boundary. These areas effectively are a form of missing data, and therefore demand special treatment, as we discuss subsequently.

An example from Muncie, Indiana, illustrates some typical changes in 1970 and 1980 tract boundaries as well as some of the potential benefits and limitations of using GIRAS data to guide interpolation (Fig. 1). Assuming that the population density in residential areas is substantially greater than in agricultural areas, then the GIRAS data indicate that we should allocate most of 1980 Tract 9.01's population to 1970 Tract 9, where most of Tract 9.01's residential area is located, while allocating relatively little population to Tract 9.01's mostly agricultural intersection with 1970 Tract 18. Similarly, if we also assume that residential areas are denser than industrial areas, we should allocate most of 1980 Tract 12's population to 1970 Tract 12, where most of its residential area lies, while allocating much smaller portions of its population to its principally industrial and agricultural intersections with 1970 Tracts 20 and 11. As long as the LULC zones are not grossly inaccurate, guiding interpolation in this manner should improve significantly on simple areal weighting, just as prior research using LULC data repeatedly has shown.

Still, GIRAS data have significant limitations. First, as with most LULC datasets, GIRAS zones are not intended to delineate zones of uniform density in population, housing or any

other census-measured characteristic. Therefore, within any single LULC class, densities may vary greatly. In Muncie, for example, both of the two 1980 tracts encompassing the central business district (CBD) and Ball State University's main campus are classified predominantly as Commercial and Services, but their population densities are, respectively, 1,498 and 5,307 persons per km². We also have identified several instances of classification error in GIRAS data, most noticeably on the outskirts of a few cities where tract data indicate that some large polygons classified as forest or rangeland have exceptionally high densities and probably are residential zones. We leave corrections of such cases for future work because a thorough quality check is a major undertaking.

A less extreme but more pervasive problem is that GIRAS data have limited spatial precision. Polygons for the urban and a few other classes have a minimum size of four hectares, whereas all other classes have a minimum size of 16 hectares, and any polygons for highways or rivers must have a minimum width of 92 meters (U.S. Geological Survey 1990). Therefore, many instances of LULC classes are over-bounded mixtures that include small instances of different land uses or covers. We also have identified occasional variations in classification "style," such that different compilation methods produce noticeable differences in the granularity of polygons and mixture of classes from one side of a tile boundary to another. As is the case for most settings that overlay spatial data from different sources, numerous boundary alignment issues also exist. Fig. 1 illustrates some relatively minor instances of this problem: several GIRAS polygon boundaries are noticeably out of alignment with tract boundaries even though the two most likely are intended to follow the same features. In other instances, boundaries are separated by 200 meters or more. Such discrepancies are of special concern in instances where a tract boundary change has occurred near a boundary between two GIRAS zones with two very different expected densities. For example, a tract boundary change may appear to occur entirely over water, in which case we might expect that no population lived in the area, although possibly a correct boundary alignment would indicate that the boundary change occurred entirely within a residential zone.

This range of issues imposes some natural limits on how effectively GIRAS data can serve as an indicator of population or housing distributions, no matter how sophisticated the model of LULC class densities is. One benefit of using geographical weighting in interpolation is that it may help to minimize many of the negative impacts of spatially imprecise and inconsistent control data. Many of the variations in the spatial precision, classification accuracy, or typical densities of LULC zones will be accentuated in certain regions. For example, commercial zones may be densely populated in some regions, and nearly unpopulated in others. Forest polygons may properly indicate low-density populations in most regions, but in others, boundary imprecision or misclassification could cause large, densely developed strips of land to be classified as forest. Geographical weighting ensures that such local anomalies influence only nearby, rather than all, estimates.

Methods

In this section, we first introduce the EM algorithm and provide its specification for our setting. We then discuss alternative density estimation approaches, identifying the potential advantages of geographical weighting and key limitations of standard linear models. We go on to specify the GWEM algorithm for areal interpolation. Finally, we discuss a potential source of large GWEM errors (density sinks), and we describe the capping procedure we use to prevent them.

The EM algorithm

The EM algorithm provides a robust framework for model fitting and maximum likelihood estimation in settings of incomplete data. Its name, coined by Dempster, Laird, and Rubin (1977) in the first general introduction to the technique, refers to the two steps that comprise each iteration of the algorithm. First, the expectation (E) step “completes” the data by computing the conditional expectation for missing data, given a set of observed data and estimated model parameters. The maximization (M) step then fits the model, estimating model parameters by maximum likelihood given the “complete” data from the E step. Thus, estimates from the E step are used in the M step to estimate a separate set of data, which are, in turn, used in the next E step to update the first estimate set, and so on, forming a feedback loop that repeats until convergence.

Flowerdew and Green (1994) demonstrate how the EM algorithm is applicable in a range of areal interpolation settings. Given a variable of interest, Y , with known values for source zones indexed by s , y_s , the aim is to estimate y_t , the unknown values of Y for target zones indexed by t . Flowerdew and Green's general approach is first to estimate y_{st} , the unknown values of Y for the regions of intersection between the source and target zones, by treating these values as missing data in an EM algorithm constrained by the known values of y_s . Then estimating y_t typically becomes a simple matter of aggregating the estimates of y_{st} for each target zone. The most suitable specification of the algorithm varies depending on the characteristics of Y , on the characteristics of ancillary data, and on the expected relationship between the two. In the simplest setting, first examined by Green (1989), Y is a count assumed to have a Poisson distribution, and the ancillary data are target zone observations of a single categorical variable. For this setting, Green derives a straightforward EM specification that essentially consists of two equations—one for each EM step.

The present application setting differs from Green's in that the ancillary control zones here are not the target zones. Nevertheless, Green's EM specification still is applicable, requiring only a substitution of the control zones for the target zones in the specification, with each “control zone” here referring to the set of all polygons sharing a single LULC class. Accordingly, the E step does not estimate values of y_{st} , as in Green's specification, but instead estimates all y_{sc} , the counts of Y for the intersections between each source zone s and control zone c , using the equation

$$\hat{y}_{sc} = y_s \left(\frac{\hat{\lambda}_c A_{sc}}{\sum_k \hat{\lambda}_k A_{sk}} \right) \quad (1)$$

where $\hat{y}_{sc}$ is an estimate of y_{sc} , $\hat{\lambda}_c$ is the estimated density of control zone c , A_{sc} is the area of the region of intersection between zones s and c (which is zero if s and c do not intersect), and k is a second control zone index, independent of c . This equation is based on the assumption that the y_{sc} have Poisson distributions with means μ_{sc} modeled as

$$\mu_{sc} = \lambda_c A_{sc} \quad (2)$$

such that the expected count of Y in any intersection between source and control zones is, intuitively, the product of the control-zone density and the area of the intersection. Given this assumption, as noted by Bloom, Pedler, and Wragg (1996), the conditional distribution of y_{sc} given y_s is a binomial distribution where the number of trials n is equal to y_s , and the probability p of a trial (an instance of Y) falling in control zone c is equal to the expected count of μ_{sc} divided by the sum of expected counts of μ_{sc} for all control zones k intersecting s . Therefore, the E step computes the conditional expectation of y_{sc} to be the expected mean of this binomial distribution, which is the product of n and p . Notably, the resultant equation

(1) has the same basic form as the general formula for dasymetric modeling provided by Mennis (2009).

To begin the first iteration of the EM algorithm, we may assume that all control zone densities are equal, in which case the first E step is equivalent to areal weighting. After that, the M step re-estimates all λ_c by fitting the model of $\hat{\mu}_{sc}$ using the estimates $\hat{y}_{sc}$ as data, which in this setting is equivalent to assigning

$$\hat{\lambda}_c = \sum_s \hat{y}_{sc} / A_c \quad (3)$$

where the summation is for all source zones s , such that the estimated density of control zone c is, intuitively, equal to the total estimated count of Y in c divided by the area of c . The estimates of $\hat{\lambda}_c$ from this equation are used to estimate $\hat{y}_{sc}$ in the next E step, which is followed by another M step, and so on. After a finite number of iterations, the changes in $\hat{y}_{sc}$ estimates between iterations of the E step become insignificantly small, and the estimates may be considered final.

All that remains is to produce the target zone estimates, which, if we assume a uniform spatial distribution of Y within each intersection between a source and control zone, are obtained by areal weighting of the estimates $\hat{y}_{sc}$:

$$\hat{y}_t = \sum_s \sum_c \frac{A_{tsc} \hat{y}_{sc}}{A_{sc}} \quad (4)$$

where A_{tsc} is the area of the intersection between target zone t , source zone s , and control zone c .

Alternatives

Although the M step of Green's EM specification produces a global estimate of density for each control zone class, the final class density estimates are localized because the E step scales estimates differently for each source zone. For example, the E step properly scales the estimated population of the commercial zone containing Muncie's CBD to have a lower density than the commercial zone containing Ball State University's main campus because the source tracts in which these two zones lie impose different population constraints. However, the E step's scaling does not alter the ratios between estimated control zone densities within source zones. If the estimated density of residential areas is twice that of commercial areas, then this ratio holds within any source zone, even if regional variations in land use or in classification procedures result in commercial areas tending to be denser than residential areas in some regions.

To address such spatial non-stationarity, one approach is to divide the study area into regions and estimate densities separately for each by refitting a linear regression model (Yuan, Smith, and Limp 1997, Langford 2006) or by sampling from source zones associated with each control zone in different regions (Mennis 2009). Similarly, the EM algorithm could be applied to different regions separately. But as Langford (2006) notes, for any regional estimation procedure, the most straightforward region definition strategies (e.g., using higher-level administrative units) are arbitrary, and discrete regionalization is questionable anyway because spatial variation in expected densities may be smoothly continuous. Langford concludes that a more suitable approach is to use GWR, as Lo (2008) and Lin, Cromley, and Zhang (2011) have done. GWR involves refitting a regression model separately for any number of locations, in each case weighting data from nearby locations more heavily than distant data (Fotheringham, Brundson, and Charlton 2002). When

employed to estimate control zone densities, this approach allows the ratios between estimated densities to vary continuously, with estimates at any location being influenced most by characteristics of the closest zones.

By permitting continuous spatial variability in density ratios, GWR may outperform global or regional regression approaches in many settings of areal interpolation, but all of these approaches, when employing linear regression as has been typical, share the same basic disadvantages; they do not inherently maintain the pycnophylactic property, and they may produce negative estimates of density and population. To address the first of these problems, a common approach is to use regression for only the initial estimates of control zone densities or subzone counts. Then, mirroring the E step given by equation (1), estimates are rescaled for each source zone (e.g., Langford 2006; Reibel and Agrawal 2007; Lin, Cromley, and Zhang 2011). This two-step procedure successfully imposes the pycnophylactic constraint, but such rescaling has a questionable meaning and effect if any of the estimated subzone counts are negative, in which case rescaling can make such estimates even more strongly negative.

To address the problem of negative estimates, Yuan, Smith, and Limp (1997) somewhat arbitrarily add to each class density estimate a value equal to the lowest negative density coefficient, raising all density estimates to ensure none are negative. Goodchild, Anselin, and Deichmann (1993) identify a number of alternatives, including lognormal regression, Bayesian modeling and—most simply—imposing a zero or small positive density coefficient in place of any negative coefficients, and then refitting the remaining positive coefficients. This last option resembles a commonly taken approach in which LULC classes expected to be sparsely populated are assigned zero densities and omitted from a model (e.g. Langford, Maguire, and Unwin 1991; Reibel and Agrawal 2007; Lin, Cromley, and Zhang 2011). However, when applying GWR, omitting LULC classes *a priori* misses the opportunity to determine whether such classes have significantly positive densities in some localities. In our applications of linear regression and GWR (for comparison with EM and GWEM approaches), we follow Goodchild, Anselin, and Deichmann's (1993) approach of omitting classes only if they yield negative coefficient estimates when included. For GWR, we handle class omissions separately for each application of weighted regression, refitting any regression as needed until all coefficients are positive at all locations.

The most robust way to avoid infeasible negative estimates is to specify a model that makes proper assumptions about data distributions. Any simple linear model relating populations to class areas is inappropriate given that these variables and the density relationships between them are strictly non-negative and strongly skewed. Therefore, the alternative nonlinear models suggested by Goodchild, Anselin, and Deichmann (1993) deserve further attention, as does the quantile regression approach that Cromley, Hanink, and Bentley (2012) use to achieve properly constrained areal interpolation. Poisson regression is another reasonable alternative for interpolating population counts, even if population distributions do not strictly conform to the independence assumption of Poisson models (Flowerdew and Green 1989). Using a Poisson generalized linear model with an identity link function generally maintains the desired non-negative constraint on estimated counts and control zone densities. Fitting a Poisson model with the EM algorithm additionally enables the incorporation of the pycnophylactic constraint.

Griffith (2010) shows that the estimates produced by an EM algorithm for a constrained linear model also may be obtained by extending a regression model to include missing-value indicator variables. Perhaps a similar strategy could be used to constrain a Poisson model. However, we choose to use the EM algorithm in this study because past research provides clear and thorough specifications of the algorithm for areal interpolation (Green 1989,

Flowerdew and Green 1994, Bloom, Pedler, and Wragg 1996), and demonstrates its suitability for interpolating historical census data (Gregory 2002; Gregory and Ell 2006). We believe its only salient limitation in our setting is the global nature of its estimated class density ratios.

The GWEM algorithm

To employ geographical weighting in the EM algorithm for areal interpolation, we execute the algorithm separately for each source zone in order to base count estimates within each zone on a unique set of local density estimates. This framework mirrors one of the GWR approaches that Lin, Cromley, and Zhang (2011) use: fitting a regression model separately for each source zone. In practice, the algorithm needs to be executed only once for each source zone that does not nest within a target zone—i.e., only where disaggregation is required; nevertheless, all source zones are included in the sampling frame for density estimates.

Beginning with the preceding EM specification, we modify the E step by replacing global density estimates with local estimates, which are “local” with respect to a principal source zone i :

$$\hat{y}_{sc}(i) = y_s \left(\frac{\hat{\lambda}_c(i) A_{sc}}{\sum_k \hat{\lambda}_k(i) A_{sk}} \right) \quad (5)$$

where $\hat{\lambda}_c(i)$ and $\hat{\lambda}_k(i)$ are local estimates of the densities of control zones c and k in the vicinity of i , and $\hat{y}_{sc}(i)$ is the estimated count of y_{sc} given local density estimates in the vicinity of i . Note that s and i are independent source zone indices, which means that this equation uses local density estimates for one source zone i to estimate counts of subzones within *any* source zone s . The reason for this peculiar feature is that for any single execution of the GWEM algorithm, the end goal is to estimate y_{ic} (the counts for subzones within the principal source zone i). Doing so requires local estimates of control zone densities, which in turn requires the $\hat{y}_{sc}(i)$ estimates of subzone counts in nearby source zones.

Specifically, local density estimates are computed in the M step using the equation

$$\hat{\lambda}_c(i) = \frac{\sum_s w_{is} \hat{y}_{sc}(i)}{\sum_s w_{is} A_{sc}} \quad (6)$$

where w_{is} is a weight determined by the distance between source zones i and s , which we measure using the zone centroids. If the w_{is} were set equal for all pairs of source zones, then this equation would be equivalent to the global M step given by equation (3). Instead, to obtain local density estimates, the weights must decrease as the distance d_{is} between i and s increases. To achieve this, we use the bi-square weighting function that commonly is used in GWR (Fotheringham, Brundson, and Charlton 2002):

$$w_{is} = \begin{cases} \left[1 - (d_{is}/b)^2 \right]^2 & \text{if } d_{is} < b \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Equation (7) produces weights that vary smoothly in a near-Gaussian s-curve from 1, at the centroid of i , to 0, at distance b (the bandwidth) and beyond. Consequently, for a single execution of the GWEM algorithm, only the source zones within distance b of i affect computations, and all others may be omitted, which can help to reduce computation time.

For each execution of the algorithm, we set b to be the distance between source zone i and the N th nearest zone, such that all executions of the algorithm base estimates on exactly N sampled source zones. Accordingly, bandwidths tend to be smaller in large cities where tracts are small and numerous, and larger in rural areas where tracts are large and more dispersed (Fig. 2). Also, because the U.S. Census Bureau released tract data only for a select set of (typically metropolitan) counties in 1980, bandwidths in some remote areas may be very large in order to span gaps in tract coverage.

To select N in our implementations of both GWR and GWEM, we re-apply areal interpolation several times using different values of N , selecting trial values with a Golden Section search algorithm, as suggested by Fotheringham, Brundson, and Charlton (2002). We then set N to be the value that yields the minimum sum of squared *interpolation errors*. This criterion differs from those suggested by Fotheringham et al. (2002), which instead involve some measure of deviation between estimates and original data for observation units. Such deviation is eliminated in the case of areal interpolation by pycnophylactic rescaling; areas yielding large regression residuals can still produce small interpolation errors. This would be the case, for example, where estimated control zone densities are initially too high or too low, but the estimated ratios between them are accurate, which is the crucial property after rescaling. Therefore, selecting N to minimize errors in the final, interpolated estimates seems most appropriate.

Density sinks and caps

A special problem case for GWEM may arise when, in the vicinity of a source zone i , a control zone c has a poor sample, such that c comprises only a small portion of a few source zones. If one of the few local instances of c occurs in a source zone with a higher-than-expected density, then zone c may act as a “density sink,” estimated to have an unrealistically high density to counterbalance underestimates in other parts of the source zone. For example, consider a region where only a few small instances of wetland exist, and one instance lies in a tract with a high density. If the tract's composition of non-wetland LULC classes indicates a low density, then GWEM could erroneously allocate most of the tract's population to the small wetland area because the local sample of wetland instances is too small to restrain the overestimation.

To prevent such occurrences, we impose caps on estimated densities after each M step. Table 1 lists the two sets of caps that we consider, each expressed as percentages of a “benchmark density,” which we set to equal the local estimate of density for the combined areas of the three typically densest classes: Residential, Small Commercial, and Mixed Urban or Built-up Land. Using relative rather than absolute caps means that class densities have higher caps in a dense city center than in a dispersed suburban area. Because our approach to selecting caps is subjective, we test two different sets of caps to investigate how sensitive results are to different selections. Set 1 reflects lower trust in GIRAS accuracy, allowing higher caps for typically low-density classes under the assumption that such areas occasionally may have high densities due to poor classification. Set 2 reflects greater trust, imposing more restrictive caps on low-density areas, including zero caps (allocating no population) for three classes. In both sets, we impose no density cap for the Residential class (because these areas generally are expected to have the highest densities) or the Large Commercial class (because these areas are large by definition, and therefore generally not at risk of being “density sinks”). The “No Data” class receives a unique treatment reflecting its unknown character. We do not “cap” its density; rather, when it occurs in the E step, we assign it a density exactly equal to the lesser of the source tract's density or the local benchmark density.

When employing density caps, the optimal bandwidth tends to be smaller because capping can reduce error caused by density sinks without increasing sample size. We apply optimization using the set of 4,280 1980 tract pairs subsequently described. When caps are not used, the optimal bandwidth is the distance to the 217th nearest neighbor. Using either set of caps, it is the distance to the 73rd nearest neighbor (the distance illustrated in Fig. 2). The optimum for “positive-coefficient” GWR (with classes that yield negative coefficients omitted) is the distance to the 126th nearest neighbor, which is between the two optima for GWEM.

Results

In this section, we first summarize and compare LULC class density estimates produced with GWEM and three other methods. We then describe the test setting we use to measure interpolation accuracy and summarize the results for several techniques. Finally, we specify and assess two hybrid techniques that combine GWEM and TDW estimates through a weighted average.

Class density estimates

To assess the GWEM algorithm, we first apply GWEM with density caps (cap set 2) to estimate local population densities of LULC classes for each of the 42,832 1980 tracts that are well covered by GIRAS tiles. Mapping the ratios between estimated densities for different pairs of LULC classes, as in Figs. 3 and 4, indicates a high degree of spatial variability in density ratios. For example, in many areas, Small Commercial densities are estimated to be more than twice the Residential densities—as in the vicinity of Bloomington, Indiana, perhaps due to campus population at Indiana University—while in many other areas, Small Commercial densities are less than a quarter of the Residential densities (Fig. 3). Ratios between estimates of Agricultural Land and Residential densities have an even broader range (Fig. 4). In northern New Mexico and north of New York City, Agricultural Land is estimated to be very sparsely populated, with estimated densities often less than a millionth of Residential densities, whereas in many Indiana tracts, Agricultural Land is estimated to have more than 1/80th the density of Residential areas. Such strong variability, even if not altogether accurate, suggests that using local rather than global estimates of density ratios should significantly improve interpolated estimates.

To gauge how GWEM deviates from other techniques, we compare its estimated densities with estimates from three other methods (Table 2): “positive-coefficient least-squares” (PCLS) regression (omitting classes that yield negative coefficients), the “global” EM algorithm, and “positive-coefficient GWR” (PCGWR). We apply each method using the same set of GIRAS-covered 1980 tracts as source zones. We employ pycnophylactic rescaling to adjust both the PCLS and PCGWR estimates, and we use no intercept term in the regression models (such that a zone with no area in any of the “positive-coefficient” classes has an expected population of zero). The final PCLS model includes only seven LULC classes, those with non-zero density estimates in Table 2.

Results reported here demonstrate that the choice of technique has a substantial impact on class density estimates. Several large shifts occur between PCLS and EM, including a near doubling in estimated Residential densities. The EM results appear more plausible, mainly because PCLS's strict assignment of zero density to several classes is counterintuitive. Many instances may have existed of 1980 population living in zones classified as, for example, Industrial, Transitional Areas, and Forest. However, compared to both sets of global estimates, the geographically weighted results seem more realistic. For example, unlike the global estimates, the local Residential density estimates generally are higher than the local Small Commercial and Mixed Urban estimates, suggesting that the global estimates for

Small Commercial and Mixed Use are biased by some exceptionally dense areas. In addition, the differences between the medians and means for the geographically weighted estimates reflect the spatial variations that we expect going from remote rural areas, where densities of most non-residential classes are likely to be near zero, to dense urban centers, where a single imprecise zone could result in a high density for any class. Comparing the two geographically weighted approaches, the salient difference is the same as for the global models: the EM approach (GWEM) desirably generates more small density estimates where the linear regression approach (PCGWR) instead assigns zeroes. Also, due to the caps used in this GWEM application, the mean GWEM estimates for several non-residential classes are more realistically lower than the PCGWR means.

Accuracy assessment

We assess accuracy by applying interpolation in a proxy setting where errors are measurable, using aggregated 1980 tract pairs as source zones and individual 1980 tracts within each pair as target zones. We begin with the 4,280 1980 tracts that, according to NHGIS tract boundaries (Minnesota Population Center 2011), do not nest within 1970 tracts (i.e., those that would require interpolation in order to produce 1980 estimates for 1970 tracts), and we merge each with its nearest neighbor. We then estimate the population of each original non-nesting tract through areal interpolation of the tract-pair populations, and compute errors by comparing estimates to the known 1980 tract data.

We initially apply nine areal interpolation techniques for comparison: areal weighting; a simulation of TDW; binary dasymetric interpolation; PCLS regression with pycnophylactic rescaling; the EM algorithm; PCGWR with pycnophylactic rescaling; GWEM without density caps; and, two applications of GWEM with different density caps (sets 1 and 2 from Table 1).

Applying true “target-density” weighting in the test setting means using 1980 tract densities to guide the interpolation of 1980 tract-pair populations, which results in 100% accuracy. Therefore, to mirror the true application setting, we simulate TDW by using 1970 tract densities to guide the interpolation of the 1980 populations, as would be the case if using TDW to estimate 1980 populations of 1970 tracts. This approach requires that each 1980 tract pair is completely covered by 1970 tracts. Therefore we omit 13 tract pairs that are not completely covered (due to changes in coverage between censuses), resulting in a final sample set of 4,267 tract pairs for all techniques.

The binary dasymetric interpolation technique is based on the simpler of two dasymetric mapping approaches specified by Wright (1936). It requires that a study region be subdivided into two control zones: populated and unpopulated areas. Then the technique essentially applies a constrained areal weighting, allocating source zone counts to target zones in proportion to the “populated” areas of intersections between them. In our implementation, we set the GIRAS classes of Residential, Commercial, and Mixed Urban or Built-up Land to be the populated areas, and all other classes to be unpopulated. This approach is perhaps the simplest means of employing LULC data to guide interpolation; yet it has performed well relative to more sophisticated techniques in several prior studies (Fisher and Langford 1995; Langford 2006; Lin, Cromley, and Zhang 2011).

In our applications of PCGWR and GWEM, when executing linear regression or the EM algorithm to estimate local control zone densities for any given tract pair, we treat that tract pair as one whole observation unit, and the component tracts of any neighboring tract pairs as separate observation units. We compute weights according to distances from the centroid of each tract pair.

Table 3 summarizes the errors produced by each of the interpolation techniques. The rightmost column reports the percentage of population estimates that deviate from the actual population by a factor greater than two (more than double or less than half) as an indication of the frequency with which each technique produces extreme proportional errors. To interpret the magnitudes of the other statistics, note that the mean population among all target tracts is 4,227. However, caution must be exercised in assessing the magnitudes because errors in the true application setting should be significantly smaller and less frequent than in this proxy setting. When producing 1980 statistics for 1970 tracts, only about 18% of all 1970 tracts require interpolation of 1980 data due to boundary changes, so most tracts are unaffected by interpolation error. Also, allocating a tract pair's population to its component tracts is a more severe form of disaggregation than what is required for most of the smaller boundary changes that actually occur between 1970 and 1980 tracts, and interpolation errors tend to be proportional to the size of boundary changes (Schroeder 2007). Still, the basic parameters of the two settings are similar enough that we expect the *relative* accuracy of different interpolation techniques to be similar in both settings. Therefore, we believe this proxy setting to be appropriate for choosing among techniques, but not (unfortunately) for evaluating the overall accuracy of any technique in a true application setting.

As expected, areal weighting produces the largest errors by far. In comparison, variations in accuracy among the remaining techniques are much smaller. Among those using LULC data, the simplest—the binary dasymetric technique—generally outperforms both of the global statistical approaches (PCLS and EM). It is weaker than the EM approach only in its slightly higher frequency of large errors (according to both the 95th percentile absolute error and the rate of large proportional errors). This outcome is in keeping with prior research that also found the binary approach to perform as well or better than global regression, even when employing pycnophylactic scaling to constrain regression-based estimates (Langford 2006; Lin, Cromley, and Zhang 2011; Cromley, Hanink, and Bentley 2012). Apparently, for most of the tract pairs, simply assuming that population is evenly distributed throughout Residential, Commercial, and Mixed Urban areas is better than assuming that population is distributed in proportion to globally estimated class densities, no matter whether those estimates were produced using linear regression or the EM algorithm. One possible factor contributing to this result may be that the globally estimated class densities are too heavily influenced by exceptional cases or regions, a condition that geographical weighting helps to correct; the geographically weighted approaches generally outperform the binary dasymetric approach.

Among the statistical approaches, the results meet our expectations that using both the EM algorithm and geographical weighting are beneficial. The global EM algorithm yields a root mean square error (RMSE) that is 7.0% less than the PCLS regression's, and PCGWR yields an RMSE that is 12.5% less than the PCLS regression's. By combining features of both the EM and GWR approaches, GWEM (when using no density caps) achieves an RMSE that is 16.4% below the PCLS regression's. Adding density caps (using cap set 2, which is slightly more effective overall than set 1) reduces GWEM's RMSE to 17.8% below the PCLS regression's. The small difference in results for the two capped GWEM applications also suggests that using different approaches to select caps, within a reasonable range, does not greatly affect results, so adopting a more rigorous cap selection strategy may be of little value. Altogether, capped GWEM using either cap set is the most accurate of the LULC-based interpolation techniques, including the binary dasymetric approach, according to all error measures.

The effectiveness of GWEM relative to TDW is not as clear-cut. All of the LULC-based techniques produce lower RMSE values than does simulated TDW, and the best GWEM

RMSE is 18.5% lower. But the mean absolute error for the best GWEM application is nearly identical to simulated TDW's, and TDW produces the lowest median absolute error of all of the initially tested techniques. The problem for TDW is that it produces an error distribution with long tails, as evidenced by its 95th-percentile absolute error being more than seven times larger than its median, whereas other techniques have a 95th-percentile that is only around four times larger. This outcome indicates that 1970 tract densities are usually a strong indicator of 1980 population distributions within 1980 tract pairs, but in a small but significant portion of cases, the 1970 tract densities differ greatly. Such marked differences occur where population distributions changed substantially between censuses, or where the part of a 1970 tract lying within a 1980 tract pair has a 1970 population density very different from the rest of the 1970 tract, making the 1970 tract's density a poor indicator of conditions within the 1980 tract pair. The large errors that simulated TDW produces in such cases give it an RMSE much greater than capped GWEM's, even though simulated TDW is more accurate in 58% of the cases.

GWEM-TDW hybrids

It seems that both GWEM and TDW—and both of the ancillary data types they employ—have unique advantages, and an optimal approach would make use of both, just as GWEM combines advantages of geographical weighting and the EM algorithm. GIRAS LULC data and 1970 tract densities could be combined into a single estimation model in many ways. To conclude the present research summary, we consider only the relatively simple possibility of combining GWEM and TDW through a weighted average. Specifically, we compute for each target tract

$$\hat{y}_H = w_G \hat{y}_G + (1 - w_G) \hat{y}_T \quad (8)$$

where $\hat{y}_G$ and $\hat{y}_T$ are population estimates for the tract given by GWEM using cap set 2 and by TDW, $\hat{y}_H$ is the “hybrid” population estimate, and w_G is a weight between 0 and 1 assigned to the GWEM estimate. The weight assigned to the TDW estimate is $(1 - w_G)$, ensuring that the two weights sum to 1, which maintains the pycnophylactic constraint as long as w_G is constant among target zones within each source zone. To find the value of that minimizes the RMSE of $\hat{y}_H$ in the proxy setting, we use ordinary least-squares (OLS) regression to fit the model

$$(\hat{y}_H - \hat{y}_T) = w_G (\hat{y}_G - \hat{y}_T) \quad (9)$$

which is derived from equation (8). The data used to fit the model are the known target tract populations and the TDW and GWEM estimates from the proxy setting. This implementation gives w_G a value of 0.633, in effect assigning 1.723 times more weight to GWEM estimates than to TDW estimates.

Results for this approach, listed as “GWEM-TDW hybrid, constant weight” in Table 3, demonstrate a substantial improvement over all other tested techniques, although the median absolute error is still higher than when using TDW alone. Therefore, we consider one other means of computing the weighted average, which is to set the weight according to an indicator of the suitability of TDW for different source zones. As noted previously and confirmed in prior research (Schroeder 2007), TDW errors tend to be larger in areas of significant population change. Therefore, we posit a model where the weight w_G in equation (8) varies as a linear function of the rate of change in each source zone:

$$w_{G,s} = \alpha_0 + \alpha_1 \widehat{\Delta y}_s \quad (10)$$

where $\widehat{\Delta y_s}$ is an estimate of the rate of change in between TDW's target and source years (in the present setting, 1970 and 1980), measured as a normalized rate, which divides the difference in values by their sum, and then scales it by 100. This measure is preferred because it eliminates the extreme positive skew and occasional infinite values produced by a typical rate measure, and it has been used previously as an effective indicator of interpolation uncertainty (Gregory and Ell 2006). Measuring relative rather than absolute change also makes the model more suitable for extension from proxy to application setting, where absolute changes generally are smaller. To estimate source zone populations in a target year (1970 populations of 1980 tract pairs) for a change measure, we use areal weighting of 1970 tract data. Although this weighting yields unreliable estimates of change, it is appropriate because both TDW and this areal weighting procedure assume uniform densities within 1970 tracts. In effect, the $\widehat{\Delta y_s}$ "change" measure might more properly be thought of a measure of the deviation between source data and TDW's ancillary data.

To set values for the coefficients in equation (10), we substitute its right-hand side for in equation (9) and fit the model using OLS. The resulting estimates are $\alpha_0 = 0.4338$ (the weight assigned to GWEM where $\widehat{\Delta y_s} = 0$) and $\alpha_1 = 0.0088$ (the amount by which $w_{G,s}$ increases for each unit increase in $\widehat{\Delta y_s}$). In a small number of tract pairs (76 of 4,267), this fit yields values of $w_{G,s}$ outside the range of 0 to 1, which can result in negative values of $\hat{y}_H$. Therefore we adjust all out-of-range values to equal 0 or 1. As summarized in Table 3, this variable-weight hybrid approach achieves the greatest accuracy of all tested techniques according to all error measures. Its RMSE is 34.6% below simulated TDW's, 28.7% below the binary dasymetric approach's, 34.0% below the PCLS regression's, 19.7% below the lowest GWEM RMSE, and 6.9% below the constant-weight hybrid approach's. The improvement over the constant-weight approach indicates that assigning more weight to GWEM estimates, and less to TDW estimates, is useful in source zones that experienced higher rates of growth.

Conclusions

Our research provides evidence that, when applying areal interpolation guided by LULC control zones to estimate population counts over a large, diverse region, the EM algorithm is more effective than linear regression, GWR is more effective than global models, and applying geographical weighting within the EM algorithm improves estimates even further. We also found that in the majority of test cases, TDW using 1970 tract densities to guide the interpolation of 1980 populations is more effective than methods using GIRAS LULC data, but the LULC-based methods generally produce fewer extreme errors than does TDW—especially, it seems, in areas of high population growth. This outcome leads us to propose, especially for the specific needs of the ISTADS project, a hybrid approach that employs both GWEM and TDW to estimate 1980 characteristics of 1970 tracts, combining the two techniques' estimates through a variably weighted average that assigns less weight to TDW in areas with higher estimated change rates.

The sizable improvements provided separately by the EM algorithm and geographical weighting in this setting suggest that GWEM should outperform linear regression (global or geographically weighted) and the global EM algorithm in most other settings with similar zone sizes and data types. The GWEM approach also yields an RMSE 11.2% lower than the binary dasymetric approach's, which is a greater reduction than other statistical models have achieved in similar settings (e.g., Langford 2006; Lin, Cromley, and Zhang 2011; Cromley, Hanink, and Bentley 2012). Still, this reduction comes at the cost of a considerably more complex and computationally intensive model. In this regard, our findings provide further evidence that in settings with less demanding accuracy requirements, the binary dasymetric

approach may offer an optimal combination of simplicity and effectiveness. Given that we also find TDW, another simple approach, to perform relatively well, especially when averaging its results with GWEM's, we suggest that averaging TDW and binary dasymetric estimates—in settings where both are applicable—may be an effective way to improve estimates without the added complexity of robust statistical modeling. Nevertheless, for our application setting, aiming to produce a multi-purpose data series for public use, we believe that the greater complexity of the GWEM-TDW hybrid is warranted.

One remaining concern is to determine whether GWEM and the GWEM-TDW hybrid approach are equally suitable when interpolating other census counts, such as total housing units, total children, or total households in poverty. Another concern is that some idiosyncrasies of our proxy test setting may bias results. For example, simulated TDW may achieve an unusually low median absolute error in the proxy setting because 1980 tract pair components are often very similar in extent to two 1970 tracts, which would not often occur in many application settings. Therefore, TDW's performance relative to GWEM's may be weaker in the application setting, and the optimal weighting for a hybrid approach could be quite different. One approach to validate the findings from the proxy setting would be to interpolate from 1980 tracts (not tract pairs) and compare results to high-accuracy estimates using manually digitized 1980 block data for selected areas.

Many other interpolation strategies exist that could produce properly constrained localized density estimates. For areal interpolation, Liu, Kyriakidis, and Goodchild (2008) use a kriging model to disaggregate the residuals remaining after applying regression-based interpolation, which not only takes account of spatial autocorrelation in residuals but also effectively imposes the pycnophylactic constraint. Cromley, Hanink, and Bentley (2012) employ quantile regression, which allows estimated density ratios to vary among source zones while also enforcing both non-negative coefficients and the pycnophylactic constraint in a way that is less sensitive to skewed data distributions. Several of the models considered by Goodchild, Anselin, and Deichmann (1993) also potentially could be extended to employ geographical weighting while maintaining proper data constraints. Future research providing comparative assessments of these techniques, including GWEM, would be useful.

Acknowledgments

This research was undertaken for the ISTADS project at the Minnesota Population Center, University of Minnesota. The project is supported by a grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, 1RO1 HD057929.

References

- Anderson, JR.; Hardy, EE.; Roach, JT.; Witmer, RE. Geological Survey Professional Paper. Vol. 964. U.S. Government Printing Office; Washington, D.C.: 1976. A Land Use and Land Cover Classification System for Use with Remote Sensor Data..
- Bloom LM, Pedler PJ, Wragg GE. Implementation of Enhanced Areal Interpolation Using MapInfo. *Computers and Geosciences*. 1996; 22:459–66.
- Cromley RG, Hanink DM, Bentley GC. A Quantile Regression Approach to Areal Interpolation. *Annals of the Association of American Geographers*. 2012; 102:763–77.
- Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*. 1977; 39:1–38.
- Fisher PF, Langford M. Modeling the Errors in Areal Interpolation between Zonal Systems by Monte Carlo Simulation. *Environment and Planning A*. 1995; 27:211–44.
- Flowerdew, R.; Green, M. Statistical Methods for Inference between Incompatible Zonal Systems.. In: Goodchild, MF.; Gopal, S., editors. *Accuracy of Spatial Databases*. Taylor & Francis; London: 1989. p. 239-48.

- Flowerdew, R.; Green, M. Areal Interpolation and Types of Data.. In: Fotheringham, S.; Rogerson, P., editors. *Spatial Analysis and GIS*. Taylor & Francis; London: 1994. p. 21-46.
- Fotheringham, AS.; Brundson, C.; Charlton, M. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons; Chichester, UK: 2002.
- Goodchild MF, Anselin L, Deichmann U. A Framework for the Areal Interpolation of Socioeconomic Data. *Environment and Planning A*. 1993; 25:383–97.
- Green, M. Research Report 3. North West Regional Research Laboratory, Lancaster University; 1989. *Statistical Methods for Areal Interpolation: The EM Algorithm for Count Data..*
- Gregory IN. The Accuracy of Areal Interpolation Techniques: Standardizing 19th and 20th Century Census Data to Allow Long-Term Comparisons. *Computers, Environment and Urban Systems*. 2002; 26:293–14.
- Gregory IN, Ell PS. Error-Sensitive Historical GIS: Identifying Areal Interpolation Errors in Time-Series Data. *International Journal of Geographical Information Science*. 2006; 20:135–52.
- Griffith, DA. Some Simplifications for the Expectation-Maximization (EM) Algorithm: The Linear Regression Model Case.; *InterStat*. 2010. p. 23March article 2 (<http://interstat.statjournals.net/YEAR/2010/articles/1003002.pdf>)
- Holt JB, Lo CP, Hodler TW. Dasymetric Estimation of Population Density and Areal Interpolation of Census Data. *Cartography and Geographic Information Science*. 2004; 31:103–21.
- Holt, JB.; Lu, H. Dasymetric Mapping for Population and Sociodemographic Data Redistribution.. In: Yang, X., editor. *Urban Remote Sensing: Monitoring, Synthesis and Modeling in the Urban Environment*. 1st ed.. John Wiley & Sons; Chichester, UK: 2011. p. 195-210.
- Langford, M.; Maguire, DJ.; Unwin, DJ. The Areal Interpolation Problem: Estimating Population Using Remote Sensing in a GIS Framework.. In: Masser, I.; Blakemore, M., editors. *Handling Geographical Information: Methodology and Potential Applications*. Longman; London: 1991. p. 55-77.
- Langford M. Obtaining Population Estimates in Non-Census Reporting Zones: An Evaluation of the 3-Class Dasymetric Model. *Computers, Environment and Urban Systems*. 2006; 30:161–80.
- Lo CP. Population Estimation Using Geographically Weighted Regression. *GIScience & Remote Sensing*. 2008; 45:131–48.
- Lin J, Cromley R, Zhang C. Using Geographically Weighted Regression to Solve the Areal Interpolation Problem. *Annals of GIS*. 2011; 17:1–14.
- Liu XH, Kyriakidis PC, Goodchild MF. Population-Density Estimation Using Regression and Area-to-Point Residual Kriging. *International Journal of Geographical Information Science*. 2008; 22:431–47.
- Mennis J. Dasymetric Mapping for Estimating Population in Small Areas. *Geography Compass*. 2009; 3:727–45.
- Minnesota Population Center. *National Historical Geographic Information System: Version 2.0*. University of Minnesota; Minneapolis: 2011.
- Mrozinski RD Jr. Cromley RG. Singly- and Doubly-Constrained Methods of Areal Interpolation for Vector-Based GIS. *Transactions in GIS*. 1999; 3:285–301.
- Noble P, Van Riper D, Ruggles S, Schroeder J, Hindman M. Harmonizing Disparate Data across Time and Place. *Historical Methods*. 2011; 44:79–85. [PubMed: 23847389]
- Price, CP.; Nakagaki, N.; Hitt, KJ.; Clawges, RM. Enhanced Historical Land-Use and Land-Cover Datasets of the U.S. Geological Survey.. U.S. Geological Survey Data Series 240. 2006. URL: <http://pubs.usgs.gov/ds/2006/240>
- Reibel M, Agrawal A. Areal Interpolation of Population Counts Using Pre-classified Land Cover Data. *Population Research and Policy Review*. 2007; 26:619–33.
- Reibel M, Bufalino ME. Street-Weighted Interpolation Techniques for Demographic Count Estimation in Incompatible Zone Systems. *Environment and Planning A*. 2005; 37:127–139.
- Schroeder JP. Target-Density Weighting Interpolation and Uncertainty Evaluation for Temporal Analysis of Census Data. *Geographic Analysis*. 2007; 39:311–35.
- Tapp A. Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources. *Cartography and Geographic Information Science*. 2010; 37:215–28.

- Tobler WR. Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association*. 1979; 74:519–35. [PubMed: 12310706]
- U.S. Geological Survey. Data Users Guide 4. U.S. Geological Survey; Reston, Virginia: 1990. Land Use and Land Cover Digital Data from 1:250,000- and 1:100,000-Scale Maps..
- Wright JK. A Method of Mapping Densities of Population with Cape Cod as an Example. *Geographical Review*. 1936; 26:103–10.
- Yuan Y, Smith RM, Limp WF. Remodeling Census Population with Spatial Information from Landsat TM Imagery. *Computers, Environment and Urban Systems*. 1997; 21:245–58.
- Xie Y. The Overlaid Network Algorithms for Areal Interpolation Problem. *Computers, Environment, and Urban Systems*. 1995; 19:287–306.

Figure 1.
An example of GIRAS-based LULC data and misaligned 1970 and 1980 tract boundaries in Muncie, Indiana. The eight LULC classes are a subset of the complete study set listed in Table 1.

Figure 2.

Bandwidths used in the application of capped GWEM, illustrated for the 1980 census tracts in three example regions: (a) northern New Mexico, (b) central Indiana, and (c) the New York City area. The range of 1.85 to 475 km in the legend is for all U.S. tracts.

Figure 3.

The ratio between the local population densities of Small Commercial and Residential zones estimated by capped GWEM for each 1980 tract in three example regions: (a) northern New Mexico, (b) central Indiana, and (c) the New York City area. The “N/A” areas have no tract data, no population, or no Small Commercial or Residential zones nearby.

Figure 4.

The ratio between the local population densities of Agricultural Land and Residential zones estimated by capped GWEM for each 1980 tract in three example regions: (a) northern New Mexico, (b) central Indiana, and (c) the New York City area. The “N/A” areas have no tract data, no population, or no Agricultural or Residential zones nearby.

Table 1

The GIRAS-Based LULC Classes and Their Density Caps

Class	Cap Set 1 (%)	Cap Set 2 (%)
Residential	<i>-none-</i>	<i>-none-</i>
Small Commercial	400	500
Large Commercial	<i>-none-</i>	<i>-none-</i>
Industrial	100	10
Transportation, Communications & Utilities	100	10
Industrial & Commercial Complexes	100	10
Mixed Urban or Built-up Land	400	500
Other Urban or Built-up Land	100	10
Agricultural Land (4)	25	2
Rangeland (3)	25	2
Forest (3)	25	2
Wetland (2)	4	0
Transitional Areas	100	100
Barren Land, Tundra & Perennial Snow or Ice (13)	4	0
Water (4)	1	0
No Data	<i>-special: see text-</i>	<i>-special: see text-</i>

Numbers in parentheses indicate how many original GIRAS classes were aggregated to form the class.

Table 2LULC Class Population Density (Persons / Km²) Estimates from Different Techniques

Class	PCLS	EM	PCGWR		GWEM, cap set 2	
			Median	Mean	Median	Mean
Residential	885	1,587	1,585	2,637	1,932	3,137
Small Commercial	2,311	2,290	1,473	1,857	1,647	2,210
Large Commercial	104	156	147	264	135	281
Industrial	0	66	47	557	81	116
Transportation, Communications & Utilities	0	36	0	273	17	116
Industrial & Commercial Complexes	583	302	228	1,044	72	93
Mixed Urban or Built-up Land	1,210	2,809	974	2,850	985	3,039
Other Urban or Built-up Land	640	336	0	374	50	96
Agricultural Land	1.8	2.3	0	18	2.7	9.4
Rangeland	0	2.3E-35	0	34	0.019	5.3
Forest	0	0.031	0	12	0.82	7.0
Wetland	0	8.0E-07	0	67	0	0
Transitional Areas	0	16	0	540	33	482
Barren Land, Tundra & Perennial Snow or Ice	0	5.0E-31	0	200	0	0
Water	0	1.0E-15	0	155	0	0
No Data	0	3.4E-07	0	28,931	402	1,292

Each row of PCGWR and GWEM statistics summarizes local class density estimates for the 1980 tracts with non-zero populations and non-zero areas in the corresponding class.

Table 3

Error Statistics for Interpolation of Populations from 1980 Tract Pairs

	Median absolute error	Mean absolute error	RMSE	95 th percentile abs. error	Pct. cases off by factor > 2
Areal weighting	1,005	1,535	2,205	4,594	16.8
Simulated TDW	390	767	1,346	2,837	6.9
Binary dasymetric	583	836	1,235	2,495	7.2
PCLS, rescaled	693	945	1,334	2,624	8.5
EM	598	855	1,241	2,445	6.9
PCGWR, rescaled, <i>N</i> =126	583	823	1,168	2,432	6.4
GWEM, no caps, <i>N</i> =217	568	788	1,115	2,255	6.1
GWEM, cap set 1, <i>N</i> =73	549	773	1,099	2,211	5.7
GWEM, cap set 2, <i>N</i> =73	546	766	1,097	2,223	5.4
GWEM-TDW hybrid, constant weight	439	639	945	1,931	4.4
GWEM-TDW hybrid, variable weight	377	581	880	1,857	4.1

Bold italics denotes the lowest value for each measure. Bold denotes the lowest values among non-hybrid approaches. *N* indicates the number of nearest neighbors used in geographically weighted approaches.