Incorporating spatial autocorrelation in dasymetric mapping: A Hierarchical Poisson Spatial Disaggregation Regression Model

Bowen He

Jonathan M. Gilligan

Janey V. Camp

Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, Tennessee, 37235

Department of Earth and Environmental Sciences, Vanderbilt University, Nashville, Tennessee, 37235

**Abstract**

**Keywords**

Population disaggregation, NLCD (National Land Cover Database), Spatial autocorrelation, Dasymetric mapping, INLA-SPDE

**Introduction**

Population distribution is one of the key inputs for investigating social-environmental coupled systems. From social vulnerability analysis to community resilience as well as environmental justice research, the importance of population spatial distribution model cannot be over stressed. Traditionally, population data are obtained through censuses survey and then aggregated up to various levels of administrative units such as block groups and tracts for reasons of privacy. Thus, detailed information on the spatial distribution of population is always unrevealed and an unrealistic impression that there are no spatial variations or non-stationarities within each administrative units is often provoked (Mennis, 2003; Wu & Murray, 2005). As a result, the modifiable areal unit problem (MAUP) that is a source of spatial statistical bias rises and severely affect the spatial data analysis results by mistakenly overestimating or underestimating the spatial data corresponding to the area or the unit of interest that is finer than the observational aggregated data (Openshaw, 1983). Thus, it is critical to develop a population spatial disaggregation model to reliably recover the spatial variation within those administrative units.

Dasymetric mapping approach is one of the cartographic techniques that recently experienced a renewed interest because of the rapid progress in computation algorithm, Geographic Information System (GIS), and remote sensing technologies (Mennis, 2009; Petrov, 2012). The technique uses information such as land cover types or other innovative ancillary data such as tax parcel data (TPD), building footprint data (BFD) and nightlight data (NLD) to redistribute observed areal aggregated data to finer spatial scales which is more suitable for various purposes like environmental justice analyses and hazard mitigation efforts evaluation since the original spatial interpretation of the demographic data in the aggregated scale is often not aligned well with the spatial scale of the hazard of interest. For instance, Mennis and Hultgren (2006) invented an intelligent dasymetric mapping (IDM) approach that combines an analyst’s subjective knowledge with an empirical sampling approach to parameterize the relationship between the ancillary National Land Cover Database (NLCD) and underlying statistical surface for purposes of reapportionment. They also identified the superiority of IDM compared to the traditional areal weighting and ‘binary’ dasymetric mapping approaches in terms of estimation accuracy (Mennis and Hultgren, 2006). Thus, the IDM was widely adopted in the environmental justice and assessment field. One example is that Giordano and Cheever (2010) used the IDM to identify communities at risk from hazardous waste generation in San Antonio, Texas, and found that people living near the generators are more often Black, or Hispanic and more likely to live below the poverty level and renters. The IDM plays an important role in localizing the spatial distribution of minorities, the poor, and non-homeowners and thus capable of identifying those socially vulnerable group to have a higher potential for exposure to larger quantity of hazardous waste generation in Bexar County, Texas than the general population (Giordano and Cheever, 2010).

Another vital dasymetric mapping technique is the Cadastral-based Expert Dasymetric System (CEDS) (Maantay et al., 2007). Rather than using uniform spatial resolution ancillary information such as the land cover data from the NLCD that is a 30m resolution, Maantay et al. (2007) developed the CEDS that uses non-uniformly distributed tax parcel information (e.g. property value, property type, and property area) to delineate the heterogeneity spatial distribution of demographic data such as population in more urban areas. The method is widely applied to solve some of the environmental and health assessment problems in more urban regions such as the New York City (NYC) (Maantay et al., 2007). Besides, the CEDS can resolve some of the problems regarding the significant underestimation of vulnerable populations as environmental hazards typically occur on a finer spatial scale than census units such as tracts or block groups. One example is that Nelson et al. (2015) successfully developed a hybrid method for creating a social vulnerability index (SoVI) at a tax parcel level by utilizing supplementary information about tax parcels to link the CEDS technique and the established social vulnerability indexing method to uncover the potential vulnerable groups that were previously masked by areal aggregated data. Nonetheless, the disadvantage of the CEDS approach is that its applicability is highly restricted by the tax parcel data quality of the study area of interest. To the best of our knowledge, unlike the NLCD database, there is no database that holds qualified tax parcel data sources across the U.S. that can be used as ancillary data for the CEDS. Thus, the CEDS procedure is regional dependent.

Other novel dasymetric mapping approaches include using built-area and height data as ancillary information (Alahmadi et al., 2014), using high-resolution address point datasets to conduct the dasymetric mapping (Zandbergen, 2011), developing multi-layer multi-class dsymetric mapping to estimate population surface (Su et al., 2010), using means of raster pixel maps to rapidly facilitate dasymetric-based population interpolation (Landford, 2007), applying the hybrid model with different ancillary data combination such as land cover data combined with tax parcel data or land cover data combined with NTL (Briggs et al., 2007; Jia and Gaughan, 2016), and using machine learning model like random forests combined with remotely-sensed ancillary data to project a finer spatial scale of demographic information distribution (Stevens et al., 2015). While these dasymetric mapping methods provide valuable insights into combining the statistical model and available ancillary data to inform the underlying finer scale distribution of demographic indicators of interest from various perspectives, none of these models incorporated the spatial non-stationarity feature. As a result, the dependent variables in those models are all treated as identically and individually distributed (iid) data, although they are spatial statistical data with autocorrelation characteristics.

Historically, several studies have been proposed to address spatial non-stationarity challenge in the population disaggregation field. For example, Li and Corcoran (2011) suggests that dividing the study area into series of subregions and performs a separate population redistribution within each subregion. The problem of this method is that the strategy of dividing the study area is arbitrary and the rather arbitrary nature of the newly divided subregions’ boundaries are unlikely to represent areas with homogeneous population distribution characteristics. Besides, local regression approaches that estimate separate coefficients for each population distribution feature are examined by quantile regression (QR) (Cromley, Hanink and Bentley, 2012) and geographically weighted regression (GWR) (Lin, Cromley and Zhang, 2011). Although these approaches advance the traditional global regression method in terms of the prediction accuracy, they still did not seem to sufficiently solve the classic problem at the population distribution’s heterogeneity and the revealing of the population spatial autocorrelation feature is highly depended on the configuration of the model’s regression covariates and their spatial distribution features (Cockx and Canters, 2015; Lee, 2011).

Thus, we aim to propose a hierarchical spatial regression model that joins a conditional autoregressive (CAR) model for the observed aggregated areal data with a Gaussian random field for the prediction sublevel grids that explicitly deals with spatial non-stationarity by incorporating two gaussian random effects to accurately reveal the spatial autocorrelation feature of the underlying population surface. Instead of spatial non-stationarity model, this proposed model aims to explain the spatial autocorrelation feature of population distribution that cannot be explained by the regression component’s structure or arbitrary dasymetric sampling strategies.

The purpose of this research is to develop a novel approach to the area-to-point spatial disaggregation problem that uses a hierarchical modeling strategy to link the areal observations and the prediction grids using these latent gaussian processes as well as the regression components so that the spatial autocorrelation pattern of population distribution can be incorporated in the dasymetric mapping model. Then, a comparative assessment of this proposed hierarchical spatial regression model (HSRM) with the IDM, traditional linear regression model (OLS), hierarchical linear regression model (HLM) and the traditional areal weighting method will be conducted by using population data in Davidson County, Nashville as a case study to test the effectiveness of this proposed model.

The remainder of this paper is structured as follows. The study area, its associated population data sets used as a case study, and the land cover covariates as ancillary data sets are discussed and summarized in Section 2. In Section 3, the IDM, OLS, HLM and the areal weighting method is briefly discussed and summarized. In addition, the proposed hierarchical spatial regression model and the accompanying Bayesian inferential process using the INLA-SPDE approach are also discussed in Section 3. In Section 4, a simulation study is implemented to examine the predictive performance of the proposed HSRM model under different scenarios. Section 5 presents the comparative assessment results of the 5 areal interpolation approaches included in this study using population data in Davidson County, Nashville as a case study. Section 6 delivers conclusions and discussions.

**Data and Study Area**

The Davidson County that is in the heart of Middle Tennessee with Nashville its county seat is used as the case study region (Figure ). It locates in 36°10′12″N, -86°46′48.00″ W with the Cumberland River flowing from east to west through the middle of the county (Figure). It encompasses 174 census tract, 487 census block groups and 9097 census blocks with a total population of 715,884, as of the 2020 decennial census (United States Census Bureau, 2021). The 2020 decennial population data for these different administrative boundaries were obtained from the decennial census data source. Like other metro city areas, the population spatial distribution in the Davidson County represents a clear autocorrelation pattern where population tends to be clustered in urban regions other than rural regions. To verify the existence of the spatial autocorrelation pattern in the population distribution, the Moran’s I analysis was conducted, and results are represented in the Results section.

Diagram

Description automatically generated with medium confidence

Figure .

The National Land Cover Database (NLCD) is used as ancillary information to serve as the fixed effects predictors in the dasymetric mapping model. The data can be accessed through the U.S. Geological Survey (USGS)’s Multi-Resolution Land Characteristics (MRLC) Consortium that is a group of federal agencies who coordinate and generate consistent and relevant land cover information at the national scale for a wide variety of environmental and land management purposes. A new generation of products named NLCD 2019 have been released that include urban imperviousness and urban imperviousness change, tree canopy and tree canopy change, and comprehensive land cover database. In this study, only the comprehensive land cover information is used to redistribute the population. For the NLCD 2019 land cover data, a nationwide data on land cover and land cover change at a 30m resolution with a 16-class legend based on a modified Anderson Level II classification system (Anderson et al., 1976) is provided. More specifically, the NLCD 2019 land cover data does not yet contain updated products for Alaska, Hawaii and Puerto Rico region and only provides information of the contiguous 48 states in the North America.

Using these population and land cover data, a series of population dasymetric maps can be created using the IDM, the OLS linear regression model, the hierarchical linear regression model, the areal weighing as well as the proposed hierarchical spatial regression model to conduct the comparative analysis and further test the effectiveness of the proposed HSRM in the context of population dasymetric mapping.

**Methods**

To evaluate the effectiveness of the proposed HSRM model, the traditional intelligent dasymetric mapping (IDM), the linear regression model (OLS), the hierarchical linear regression model (HLM), and the traditional areal weighting mapping (AWM) are tested for the purpose of comparative analysis. Here, a brief review of these approaches is provided in the context of area interpolation and dasymetric mapping using NLCD as ancillary information or model predictors. Following, the proposed HSRM model and its associated INLA-SPDE framework is elaborated.

**Areal Weighting Mapping (AWM)**

The Areal Weighting Mapping (AWM) method assigns a proportion of the population to the target area relative to the percentage of its area within the source geographic unit that it belongs to (Holifield et al., 2017). The core assumption of this traditional mapping approach is that the population is homogeneously spatially distributed within the source geographic unit like census tracts.

**Intelligent Dasymetric Mapping (IDM)**

The IDM is an ‘intelligent’ dasymetric mapping technique that uses a data-driven methodology or an analyst’s prior knowledge to specify the functional relationship between the ancillary land cover classes with the underlying statistical surface being mapped such as population surface (Mennis and Hultgren, 2006). The core of this approach is that it employs a flexible empirical sampling to acquire information on the areal densities of individual ancillary classes so that the ratio of different class densities can be used to guide to redistribute the areal spatial data to the target spatial zone areas controlled by the ancillary data resolution (Mennis and Hultgren, 2006). The first step of the IDM process is to choose the appropriate sampling strategy to link source zones with each ancillary land cover class (Mennis and Hultgren, 2006). Three sampling methods were developed: (1) ‘containment’ method: samples those source zones that are wholly contained within an individual ancillary class; (2) ‘centroid’ method: selects those source zones that have their centroids contained within an individual ancillary class; (3) ‘percent cover’ method allows the user to set a threshold percentage value and then selects those source zones whose area of occupation of a single ancillary class is equal to or exceeds that threshold (Mennis and Hultgren, 2006). Once the dasymetric sampling strategy is determined, the sampling process establishes the link between each ancillary land cover class and its corresponding source zones that can represent each of these classes based on the selected sampling strategy criterion (Mennis and Hultgren, 2006). Then, the estimated density of ancillary class can be computed:

where is the estimated density of the ancillary class , is the count or value of the sampled source zones, is the area of the sampled source zones that represent the ancillary class , is the total number of sampled source zones that represent the ancillary class .

For these ancillary classes that have been successfully sampled, their densities can be computed and the estimated count for a given target zone can thus be obtained:

where is the estimated count for a given target zone, is the area of the target zone, and is the estimated density of ancillary class that resides in the target zone .

For certain ancillary classes when no source zones can meet the sampling criterion to represent them, the “refined” areal weight is first applied to derive their temporary count:

where is the temporary estimated count of the target zone associated with unsampled ancillary class , is the estimated density of known class within the source zone , is the area of the target zone associated with known class , and is the area of the target zone associated with unknown class . Once the temporary estimated count of the unknown ancillary class is calculated for each target zones in the entire data set, the estimated density of ancillary class can thus be derived:

where is the estimated density of class and is the total number of target zones in the entire data set associated with class *.*

Based on the computation process illustrated above, several potential issues can be identified for the IDM approach. First, the IDM approach suffers from the imbalance of the inadequate sampling and sampling bias. Taking the ‘percent cover’ sampling strategy as an example, when the threshold is set low, enough number of source zones can be sampled to represent the ancillary land cover classes. Nonetheless, the quality of those sampled source zones is questionable since the threshold is low and might not be able to truly reflect the ancillary land cover class density. On the other hand, when the threshold is high, the sampled source zones that can meet the criterion are sufficient to represent the true relationship between the data density and the land cover type. However, the number of sampled qualified source zones might be very low since few source zones can meet the high criterion. As a result, many ancillary land cover classes cannot be successfully sampled and the IDM thus degenerates into refined area weighting approach.

Another issue is that the process lacks the understanding of its ‘black-box’ sampling design. For different study area of interest, the IDM sampling design can play a critical role in determining the dasymetric sampling accuracy. In other words, the characteristics of the study area and its associated ancillary data can determine the best sampling design. However, it is extremely difficult to identify the best sampling design for the study area of interest since the true underlying surface is always unavailable and the comparison is inapplicable. Thus, designing the best or appropriate sampling strategy for IDM is a challenge.

**Linear regression model (OLS)**

The linear regression model is often applied in dasymetric mapping to measure the vulnerable population during natural hazards or devising emergency evacuation strategies (Bian and Wilmot, 2015, 2017). The core difference from the IDM is that instead of applying an arbitrary dasymetric sampling method, an ordinary least squares (OLS) linear regression estimation is used to derive the coefficient of ancillary land cover class densities. In this paper, we applied three regularized regression analysis that are: (1) LASSO whose regularization term (L1 regularization norm) penalizes absolute value of the regression coefficients (, is th regression coefficient); (2) Ridge whose regularization term penalizes the square of the magnitude of the regression coefficients (); (3) Elastic Net regression whose penalization algorithm uses a weighted combination of LASSO and Ridge regularization term (( + ), where controls the weighting scheme between L1 and L2 norm. The benefits of these regularization schemes are to help automatically identify the significant land cover covariates to be included in the linear model structure.

Once the significant land cover covariates are identified, the linear regression model in the context of population dasymetric mapping can be elaborated as follows. Suppose the population that resides in each grid within census tract can be predicted based on the total population density of that census tract and the selected land cover covariates of the grid based on the linear regularized analysis. Thus, the grid population density is modeled:

where is the population count modeled at the grid , is the selected significant land cover covariates value based on the linear regularized analysis, is the land covariates coefficients, is the intercept.

However, since the grid ’s population value is unobserved, the model at grid scale can only be evaluated at the census block level since it is the finest scale of observed data, and the sum of the population at every grid within block is observed. Thus, for every census block within tract , we have a linear regression model without intercept term

where population count at block , is the vector that contains the sum of each selected land cover covariates value of all grids within the census bock , is the total number of the grids within the census block .

After the coefficient between each land cover type and population is estimated, the population in each grid can be redistributed based on the relative importance of the grids’ land cover in the source zones while remains the pycnophylactic property (volume conservation) that the sum of all the redistributed grids’ population still equals to the source zone’s original total population:

where is the estimated number of persons in the grid , is the estimated number of persons in the tract that the grid resides in, is the estimated linear regression coefficient for land cover and is the area of the grid that the land cover resides in.

**Hierarchical Linear Model (HLM)**

Hierarchical Linear Modeling (HLM) is a complex form of ordinary least squares (OLS) regression that is used to analyze variance in the outcome variables when the regression components are at varying hierarchical levels (Woltman et al., 2012). While the method is widely used in many fields including psychology (Wolman et al., 2012) and educational applications (Raudenbush, 1988), there is no previous work that applied the hierarchical linear model structure on the analysis of dasymetric mapping as far as we are aware.

In the context of applying HLM to the population dasymetric mapping problem, one core assumption is that population data shares similar variance in a higher-level data structure such as the census tract, county, or even state, and the difference among these higher-level-shared variances is significant. HLM accounts for the shared variance in hierarchically structured data by applying a “soft constraint” to the intercept or regression coefficients in the model so that they can vary from one higher-level group to another higher-level group (Gelman and Hill, 2006). Besides, the hierarchical model can extract information from the complete-pooling estimates for these groups with fewer observations (Gelman and Hill, 2006). Our hypothesis is that the hierarchical model structure can address some of the spatial non-stationarity challenge in the dasymetric redistribution process, nonetheless, it is restricted by the arbitrary nature of administrative grouping strategy since the way of grouping grids or blocks’ population into tracts or counties does not guarantee the homogeneous or constant variances of population distribution characteristics within these administrative groups.

For the sake of simplicity and data size restriction reason, we use a two-level linear regression model structure with varying intercept in the dasymetric mapping model:

, for *j* = 1,….,*m* (First level population group, tract)*;*

*i* = 1,….,*n* (Second level population data, block)

where is the modeled population data in the grid *i* within tract *j*, is the varying intercept corresponding to the tract *j*, = {, ,…,} is the selected land cover land use predictor set identified in the previous regularized regression analysis, and is the unexplained variances for the tract *j*.

After the HLM is fitted, the grids’ population can be redistributed similar as the OLS linear regression model except that for each different tract, the intercept is different:

By incorporating a varying intercept, the relative importance of each land cover in each different tract is automatically adjusted by the hierarchical model structure so that the non-stationarity of the variances in the integral population data is not neglected.

**Hierarchical Spatial Regression Model (HSRM)**

The hierarchical spatial regression model in the context of population disaggregation is proposed as follows. Suppose denotes the study area of the interest that can be partitioned into areal units ,…., and ,…..,are their corresponding areal population linear predictor observations. Suppose denotes the number of persons at grid point within area that is unobserved. The purpose is to predict the quantity of interest over a set of finer grid points ,…., from coarser areal observations. Two levels of spatial random effects are incorporated to characterize the spatial autocorrelation features at both grids level and the areal level. The proposed population HSRM is characterized by a Poisson Spatial Regression Model

*,*

*,*

where the linear predictor for both grids and areal level with spatial random effects is characterized as follows

First level areal observations: , = 1,…,

Second level grids predictions: , = 1,…,

and are vectors of selected land cover covariates for the th area and th grid point, respectively. Both two covariates’ vectors include an intercept term, and is the corresponding regression coefficients vector. To be specific, the land cover covariate values for the th area is obtained as areal averages of all corresponding grid points values within the area , where is the size of the th area and is the land cover covariate value for grid point *s* within the th area. The spatial random effects characterize spatial autocorrelation at the grid point level in the model that assumes population in closer grids is more similar than that in grids far apart. These random effects are assumed to be a zero-mean stationary gaussian random field, that is , where is a spatially structured, positive definite variance-covariance matrix that is assumed to follow the Matérn family of covariance functions such that for generic grid point and , we have

where denotes the Euclidean distance between the spatial grid point and , is the marginal variance of the process, is a scaling parameter that controls the range which is the distance where spatial correlation is approximately 0.1 (Matérn, 1986). is the modified Bessel function of the second kind, and is the smoothness parameter that is often set as a constant due to identifiability issues (Abramowitz and Stegun, 1972). Here, we set based on Lindgren et al. (2011).

The second set of areal level spatial random effects characterizes spatial autocorrelation in the observed areal data and are assigned a conditional autoregressive (CAR) prior. We apply the CAR model proposed by Leroux et al. (2000) which was found that this CAR model outperformed other models in a recent study (Lee, 2011). These spatial random effects are also assumed to follow a zero-mean gaussian random process that , where is a precision matrix and is the marginal variance parameter of the gaussian process. Specifically, , whereis a spatial autocorrelation parameter, is an vector of 1’s, is the identity matrix and is a binary matrix capturing the neighborhood information of the areas. For which if areas and are neighbors to each other and otherwise. is the spatial random effect of the area that the th grid belongs to.

**Bayesian inference using the INLA-SPDE approach**

The INLA-SPDE approach (Rue et al., 2009; Lindgren et al., 2011) is used to fit the model parameter and predict the missing underlying grids point population data. Suppose is the parameter vector and based on the Bayesian inference theory, its joint posterior distribution is proportional to:

where is the joint prior distribution of . The Integrated Nested Laplace Approximation (INLA) method applies the Laplace approximation method to provide a numerical approximation of the marginal posterior distribution of each model parameter element of the .

The stochastic partial differential equation (SPDE) consists in representing a continuous spatial random process using a discretely indexed spatial random process which is a Gaussian Random Markov Field (GMRF).

where , is Laplacian, controls the smoothness, is the scale parameter, controls the marginal variances of the Matérn covariance function, and is a Gaussian spatial white noise process.

The exact solution of this linear fractional SPDE is verified to be the Gaussian random field with the Matérn variance-covariance function (Blangiardo and Cameletti, 2015). Thus, the problem becomes using the finite element method through a basis function representation to approximate the exact solution to the SPDE. The basis function representation can be defined on a triangulation of the domain , given by

where G is the total number of vertices of the triangulation, { is the set of basis functions, and {} are zero mean Gaussian distributed weights.

Thus, for each linear predictor, we have

where is the value of the gth basis function evaluated in the grid point. More generally, it is possible to express the linear predictor as

where is the generic element of the sparse matrix *A* which maps the GMRF from the *G* triangulation vertices to the observational locations. For more detailed information, please see Blangiardo and Cameletti. (2015).

In summary, fitting a Matérn correlated spatial mixed-effects model in INLA consists of following several steps:

1. Create a mesh to approximate the spatial effect
2. Create a projection matrix to link the observations to the mesh
3. Set the stochastic partial differential equation (SPDE)
4. Specify the model predictions dataset that are the target grid locations
5. Put everything together in a stack object
6. Fit the model and predict the values for target grid locations

For the comparative analysis, the grids population assigned by 5 different spatial interpolation approaches elaborated above will be aggregated in each census block to compare with the true Decennial census blocks population data through computing three interpolation accuracy performance metrics: (1) R2; (2) Root Mean Squared Error (RMSE); (3) Mean Absolute Error (MAE).

The R code for the comparative and simulation analysis is provided in the online Supplemental

Materials.

**Simulation Study**

The purpose of the simulation study is to demonstrate the spatial interpolation ability of the proposed spatial model as well as its advancement compared to the traditional areal weighting method in the context of intended dasymetric mapping applications. The underlying true grids value were generated in the unit of [0,1][0,1] square as the study area with the resolution of 0.01 (e.g. 100100 grids). The observational area is set up with the resolution of 0.1 (e.g. 1010 areas) and the areal value is computed by aggregating every 100 underlying true grids value that belong to that area to obtain their area’s value. The simulation context is predicting the finer resolution underlying true grids value from the coarser observational areal value using the proposed hierarchical spatial model. Here, the simulation study is conducted in two stages. In the first stage, a single-level spatial regression model with only grids level Matérn spatial random effects () included is used to generate the underlying true grids value and interpolate the grids value from coarser areal observational values. Then, a hierarchical spatial regression model with both the grids level Matérn spatial random effects and the areal CAR spatial random effects included is tested in the simulation study. The Poisson spatial regression model in this simulation study is configured as follows:

*,*

*,*

First stage single-level: ,

Second stage hierarchical: ,

The true values of the parameters of the model used in generating the underlying true grids data are described as follows. For the first level grids Matérn spatial process, we set up the scaling parameter , marginal variance of the random process = 1, smooth parameter 𝜈 = 1. For the second level CAR spatial process, we set up the marginal variance parameter = 0.25, spatial autocorrelation parameter 𝜌 = 0.7. The intercept is set as = 2. A hundred replicate data sets were generated from the model in the simulation study.

To evaluate the predictive performance of the model and its advancement compared to the traditional areal weighting approach, we computed 3 metrics: R2, RMSE, MAE for both the two configured Poisson spatial regression models as well as the traditional areal weighting for each simulated replicate data sets.

Table

Description automatically generated with medium confidence

Figure . Plots of the (a) areal (1010); (b) grid (100100) configurations used in the simulation study. These were generated on the unit [0,1][0,1] square.

**Results**

**Spatial Autocorrelation Analysis**

The Moran’s I analysis is conducted on the tract population density for the study area and the results is exhibited in the Figure to reveal the region’s population spatial autocorrelation characteristic.

Graphical user interface, application

Description automatically generated Graphical user interface

Description automatically generated

Figure

**Areal Weighting Mapping (AWM)**

In the AWM process, population is evenly distributed to every land cover type except that the open water and herbaceous wetlands have absolutely no population. Thus, the spatial distribution of the Areal Weighting Mapping (AWM) grids population exhibits a homogeneous pattern with no variances within the census tracts. The grids population interpolated by the AWM method is shown in the Figure (d). The AWM assigned grids population in each census block is aggregated to compare with the true Decennial census blocks population data by three performance metrics: R2, RMSE, MAE. The R2 is 0.32, RMSE is 150.00, and MAE is 60.46.

**Intelligent Dasymetric Mapping (IDM)**

In the IDM process, the 0.5 percent cover sampling method is applied that those source zones whose area of occupation of a single ancillary class is equal to or exceeds 50% are selected to calculate the population density of that ancillary land cover class. The open water and herbaceous wetlands are predefined to have 0 population density. The detailed information regarding the sampled and unsampled land cover population coefficient is summarized in the Table. The grids population interpolated by the IDM method is shown in the Figure (c). Similar, the IDM assigned grids population in each census block is aggregated to compare with the true Decennial census blocks population data and R2 is 0.21, RMSE is 161.08, and MAE is 62.49.

Table

|  |  |  |  |
| --- | --- | --- | --- |
| Sampled Land Cover | | Unsampled Land Cover | |
| Type | **Density (1/m2)** | **Type** | **Density (1/m2)** |
| developed open space | 9.3E-4 | barren | 6.73E-02 |
| developed low intensity | 1.7E-3 | grassland | 5.97E-03 |
| developed medium intensity | 2.9E-3 | evergreen forest | 1.9E-03 |
| developed high intensity | 1.4E-3 | mixed forest | 8.2E-03 |
| deciduous forest | 8.83E-05 | shrub | 1.97E-02 |
| pasture | 3.3E-04 | cultivated crops | 0 |
|  |  | woody wetlands | 3.6E-03 |

**Linear Regression Model (OLS)**

For the OLS linear regression model, only the 8 land cover types that were selected in the regularized linear regression model are included in the OLS linear regression model and only 3 of them have positive coefficients. The open water and herbaceous wetlands are predefined with 0 population density. The detailed information associated with each land cover population coefficient is summarized in the Table. The grids population interpolated by the OLS mapping method is shown in the Figure (a). The OLS mapping model significantly improve the spatial interpolation accuracy compared with the IDM and AWM approach. The OLS assigned grids population in each census block is aggregated to compare with the true Decennial census blocks population data and the R2 is 0.67, RMSE is 103.79, and MAE is 44.75.

Table

|  |  |
| --- | --- |
| Land Cover | Coefficient (1/900 m2) |
| developed open space | 7.83E-02 |
| developed low intensity | 1.77 |
| developed medium intensity | 2.82 |
| developed high intensity | 0 |
| deciduous forest | 0 |
| pasture | 0 |
| barren | 0 |
| grassland | 0 |
| evergreen forest | 0 |
| mixed forest | 0 |
| shrub | 0 |
| cultivated crops | 0 |
| woody wetlands | 0 |

**Hierarchical Linear Model (HLM)**

In the HLM mapping process, the coefficient for all the 8 selected land cover types is fixed except that the developed medium intensity can vary across the tracts so that the varying relative importance of the medium intensity land across the tracts can be considered. The varying coefficients of the developed medium intensity for each tract are represented in the Figure that clearly exhibits the pattern that the medium intensity land has denser population in the urban tracts than rural tracts. The grids population interpolated by the HLM mapping method is shown in the Figure (b). The detailed information associated with each land cover population coefficient is summarized in the Table . The HLM assigned grids population in each census block is aggregated to compare with the true Decennial census blocks population data and all performance metrics show limited improvement compared to the OLS mapping model. The R2 is 0.69, RMSE is 101.10, and MAE is 43.42.

Table

|  |  |
| --- | --- |
| Land Cover | Coefficient (1/900 m2) |
| developed open space | 2.03E-01 |
| developed low intensity | 1.74 |
| developed medium intensity\* | 0.83 |
| developed high intensity | 0 |
| deciduous forest | 0 |
| pasture | 0 |
| barren | 0 |
| grassland | 0 |
| evergreen forest | 0 |
| mixed forest | 0 |
| shrub | 0 |
| cultivated crops | 0 |
| woody wetlands | 0 |

\*developed medium intensity: coefficients vary across the tracts.

Chart, radar chart

Description automatically generated

Figure Developed medium intensity coefficients that vary across tracts.

Graphical user interface, application

Description automatically generated

Figure The grids population interpolated by the 4 dasymetric mapping methods: (a) OLS mapping method; (2) HLM mapping method; (3) IDM method; (4) AWM method.

**Hierarchical Spatial Regression Model (HSRM)**

**Simulation Study**

*Single-level Spatial Regression Model*

Chart, histogram

Description automatically generated

Figure Predictive performance evaluation results of the single-level Possion spatial regression model using 100 simulated realizations.

Calendar

Description automatically generated

Figure Visualizations of 3 simulated realizations by single-level Poisson spatial regression model and its interpolation accuracy comparison with the AWM approach.

*Hierarchical Spatial Regression Model*

*Chart

Description automatically generated*

Figure Predictive performance evaluation results of the hierarchical Possion spatial regression model using 100 simulated realizations.

*Calendar

Description automatically generated*

Figure Visualizations of 3 simulated realizations by hierarchical Poisson spatial regression model and its interpolation accuracy comparison with the AWM approach.

*Davidson County Population*

The HSRM assigned grids population in each census block is aggregated to compare with the true Decennial census blocks population data and the R2 is , RMSE is and MAE is .

**Discussion**

**Reference**

Abramowitz, M. and Stegun, I. Handbook of mathematical functions. New York: Courier Dover Publications, 1972.

Alahmadi, M., Atkinson, P. M., & Martin, D. (2014). A comparison of small-area population estimation techniques using built-area and height data, Riyadh, Saudi Arabia. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(5), 1959-1969.

Anderson, James R., Ernest E. Hardy, John T. Roach, and Richard E. Witmer. 1976. A Land Use and Land Cover Classification System For Use With Remote Sensor Data. USGS Professional Paper 964. A revision of the land use classification system as presented in the USGS Circular 671.

Bian, R., & Wilmot, C. G. (2017). Measuring the vulnerability of disadvantaged populations during hurricane evacuation. Natural Hazards, 85(2), 691-707.

Bian, R., & Wilmot, C. G. (2015). Spatiotemporal population distribution method for emergency evacuation: case study of New Orleans, Louisiana. *Transportation research record*, *2532*(1), 99-106.

Blangiardo, M., & Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.

Briggs, D. J., Gulliver, J., Fecht, D., & Vienneau, D. M. (2007). Dasymetric modelling of small-area population distribution using land cover and light emissions data. Remote sensing of Environment, 108(4), 451-466.

Cockx, K., & Canters, F. (2015). Incorporating spatial non-stationarity to improve dasymetric mapping of population. *Applied Geography*, *63*, 220-230.

Cromley, R. G., Hanink, D. M., & Bentley, G. C. (2012). A quantile regression approach to areal interpolation. Annals of the Association of American Geographers, 102(4), 763-777.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

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.

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.

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.

Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. Spatial and spatio-temporal epidemiology, 2(2), 79-89.

Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In Statistical models in epidemiology, the environment, and clinical trials (pp. 179-191). Springer, New York, NY.

Li, T., & Corcoran, J. (2011). Testing dasymetric techniques to spatially disaggregate the regional population forecasts for Southeast Queensland. Journal of Spatial Science, 56(2), 203-221.

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

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4), 423-498.

Holifield, R., Chakraborty, J., & Walker, G. (Eds.). (2018). The Routledge handbook of environmental justice (pp. 1-11). London, UK: Routledge, Taylor & Francis Group.

Maantay, J. A., Maroko, A. R., & Herrmann, C. (2007). Mapping population distribution in the urban environment: The cadastral-based expert dasymetric system (CEDS). Cartography and Geographic Information Science, 34(2), 77-102.

Matérn, B. Spatial variation. Berlin: Springer, 1986.

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

Mennis, J. (2009). Dasymetric mapping for small area population estimation. Geography Compass,

3, 727–745.

Nelson, K. S., Abkowitz, M. D., & Camp, J. V. (2015). A method for creating high resolution maps of social vulnerability in the context of environmental hazards. Applied Geography, 63, 89-100.

Openshaw, S. (1983). The modifiable areal unit problem. In Concepts and techniques in modern geography (Vol. 38). Norwich: Geobooks.

Petrov, A. (2012). One hundred years of dasymetric mapping: back to the origin. The Cartographic

Journal, 49(3), 256-264.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *journal of Educational Statistics*, *13*(2), 85-116.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology), 71(2), 319-392.

United States Census Bureau. (2021). Population Estimates, July 1, 2021 (V2021) – Davidson County, TN. *QuickFacts*. Retrieved from [https://www.census.gov/quickfacts/fact/table/davidsoncountytennessee#](https://www.census.gov/quickfacts/fact/table/davidsoncountytennessee)

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely sensed and ancillary data. PloS one, 10(2), e0107042.

Su, M. D., Lin, M. C., Hsieh, H. I., Tsai, B. W., & Lin, C. H. (2010). Multi-layer multi-class dasymetric mapping to estimate population distribution. Science of the Total Environment, 408(20), 4807-4816.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in quantitative methods for psychology*, *8*(1), 52-69.

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

Zandbergen, P. A. (2011). Dasymetric mapping using high resolution address point datasets. Transactions in GIS, 15, 5-27.