Computers, Environment and Urban Systems 75 (2019) 132–145

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling

Parmanand Sinha[a,*], Andrea E. Gaughan[a], Forrest R. Stevens[a], Jeremiah J. Nieves[b,c], Alessandro Sorichetta[b,c], Andrew J. Tatem[b,c]

[a] *Department of Geography and Geosciences, University of Louisville, Louisville, KY, USA*
[b] *WorldPop, Department of Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK*
[c] *Flowminder Foundation, Stockholm, Sweden*

ARTICLE INFO

Keywords:
Dasymetric modeling
Random forest
Spatial autocorrelation
Gridded population modeling

Abstract: Gridded human population data provide a spatial denominator to identify populations at risk, quantify burdens, and inform our understanding of human-environment systems. When modeling gridded population, the information used for training the model may differ in spatial resolution than what is produced by the model prediction. This case arises when approaching population modeling from a top-down, dasymetric approach in which one redistributes coarse administrative unit level population data (i.e., source unit) to a finer scale (i.e., target unit). However, often overlooked are issues associated with the differing variance across the scale, spatial autocorrelation and bias in sampling techniques. In this study, we examine the effects of intentionally biasing our sampling from the source to target scale within the context of a weighted, dasymetric mapping approach. The weighted component is based on a Random Forest estimator, which is a non-parametric ensemble-based prediction model. We investigate issues of autocorrelation and heterogeneity in the training data using 18 different types of samples to show the variations in training, census-level (i.e., source) and output, grid-level (i.e., target) predictions. We compare results to simple random sampling and geographically stratified random sampling. Results indicate that the Random Forest model is sensitive to the spatial autocorrelation inherent in the training data, which leads to an increase in the variance of the residuals. Sample training datasets that are at a spatial scale representative of the true population produced the best fitting models. However, the true representative dataset varied in autocorrelation for both scales. More attention is needed with ensemble-based learning and spatially-heterogeneous data as underlying issues of spatial autocorrelation influence results for both the census-level and grid-level estimations.

1. Introduction

A spatially-explicit human denominator provides a foundation for identifying populations at risk, quantifying burdens, mapping dynamics of infectious disease and generally informing our understanding of human distribution and movement, both over space and time (Hay, Noor, Nelson, & Tatem, 2005; Linard, Gilbert, & Tatem, 2011; Pezzulo et al., 2016a; Tatem, 2014; Tatem et al., 2012; Tejedor-Garavito et al., 2017). While reference data for population traditionally comes from censuses or surveys (Tatem, 2014), the tabulation of the population and demographic information must then be tied to irregular and varying sized administrative units for spatial representation (Tatem et al., 2012). To improve the consistency and comparability of these data, research advances continue to improve on techniques to grid the data, creating uniform, areal units for a final raster-based population product at a given spatial grain (Azar, Engstrom, Graesser, & Comenetz, 2013; Bhaduri, Bright, Coleman, & Urban, 2007; Sorichetta et al., 2016; Stevens, Gaughan, Linard, & Tatem, 2015). Gridded population datasets have been used extensively in natural disaster operations (The National Research Council, 2007), epidemiological modeling (Linard & Tatem, 2012; Pezzulo et al., 2016b), identifying populations at risk to climate and land change (Füssel, 2007; Hahn, Riederer, & Foster, 2009; López-Carr et al., 2014; Vargo, Habeeb, & Stone, 2013), and infectious disease and hazards (Salje et al., 2016). A variety of different approaches for creating gridded population data, coupled with increasingly detailed ancillary information, underlies a direct need to better understand the spatial and, when appropriate, statistical structure driving these models and how that is affected at multiple spatial scales (Nieves et al., 2017).

[*] Corresponding author at: Room 210, Lutz Hall, University of Louisville, Louisville, KY 40292, USA.
E-mail address: parmanand.sinha@louisville.edu (P. Sinha).

https://doi.org/10.1016/j.compenvurbsys.2019.01.006
Received 3 July 2018; Received in revised form 3 December 2018; Accepted 12 January 2019
Available online 08 February 2019

The methods that generate these gridded data range from straightforward areal weighting techniques (also known as proportional reallocation) to dasymetric (weighted) redistribution, and more statistically advanced models requiring ancillary information, often in the form of remotely-sensed or other spatially-explicit data sources (Bright, Rose, & Urban, 2016; Gaughan, Stevens, Linard, Jia, & Tatem, 2013; Mennis, 2003; Mennis & Hultgren, 2006). In addition to the different methods that produce gridded datasets, it's important to note the type of population may be different as well. For instance, some products might model residential population (i.e., night-time population) (Doxsey-Whitfield et al., 2015) while others (Bhaduri et al., 2007) may focus on ambient population (i.e., the average location of people across time). A well-known dataset that uses areal weighting is the Gridded Population of the World (GPWv4), producing gridded population data at 30 arc-seconds spatial resolution (~1-kilometer resolution at the equator) (CIESIN, 2016). The process of dasymetric mapping requires disaggregating spatial data from coarser "*source*" units into finer "*target*" units using ancillary data at a given target spatial resolution (Mennis, 2003; Mennis & Hultgren, 2006). The variability in the ancillary data values (e.g., land cover) enables an asymmetric allocation of population values from the source spatial resolution (Nieves et al., 2017). For example, this happens when redistributing population in such a way that urban areas will have a higher weight than forested areas (Bhaduri, Bright, & Rose, 2014).

Out of many available methods of dasymetric mapping, the intelligent dasymetric mapping is one of the most widely used and most flexible techniques (Mennis & Hultgren, 2006; Nagle, Buttenfield, Leyk, & Spielman, 2014). This technique downscales from source populations P_s to target populations P_t as follows:

$$P_t = P_s \frac{w_t}{\sum_{t \in s} w_t} \tag{1}$$

where the numerator w_t is the expected population count in target area t and denominator is the sum of all expected counts w_t in that source area. The expected population, w could be written as a function of different covariates:

$$w = f(c_1, c_2, \ldots, c_n) + error \tag{2}$$

where c_i represents the individual covariate such as lights-at-night, slope, elevation, and proximity to land-use types. In a statistical model, the expected population density w_t might be derived through a regression type approach (Mennis, 2009) or ensemble predictor (e.g. Random Forest) using population data combined with ancillary data layers (Stevens et al., 2015). For example, the WorldPop Project (www.worldpop.org) uses a random forest (RF) statistical model (Breiman, 2001) to generate a predictive weighting layer based on a suite of ancillary data that that is then anchored through dasymetrically disaggregating census counts into a three arc second resolution (~100 m at the equator) gridded product (Stevens et al., 2015). This combined statistical and dasymetric approach has been shown to be more accurate with final gridded population datasets (Gaughan et al., 2016; Sorichetta et al., 2015; Stevens et al., 2015). However, validation at the grid (i.e. target) scale is difficult as the model is paramaterized with an aggregated set of counts (i.e. source) at the administrative unit level.

Due to a change in the model scale, the range of population density of the target data differs from the source. The difference in population density of urban vs. rural areas may impose bimodality at the finer scale distribution that may be absent at the source scale distribution. This mismatch would be wider with the increase in the size of the source administrative unit. Based on the method of modeling, the prediction based on source training samples may lead to underestimation of dispersion and extremes in the distribution at target scale.

In this context, we examined the effect of mismatch related to range, variability, and spatial structure of spatially downscaling population counts from source administrative units to target grid cells. We use spatial autocorrelation statistics as a measure of the spatial structure. At administrative unit (i.e., source) scale, we analyze the quality of prediction on a holdout dataset and explore the performance of the RF statistical model with a spatially autocorrelated dataset. At the grid scale (i.e., target) we analyze the variation in prediction using the source scale model and covariates at the target scale. The mismatch is generated by varying the level of spatial autocorrelation of samples drawn from census data. Two other sampling methods, simple random sampling (SRS) and geographically stratified random sampling (GSRS), provide a benchmark to assess sampling performance. Using the RF model, we tested the variation of predictions, errors, and spatial patterns of residuals. Lastly, we examined levels of bias introduced in an aggregated prediction at the source unit level by an RF-informed dasymetric weighting technique across the samples.

2. Methods and data

The method and datasets are described in four parts. The first part introduces the study area and the ancillary datasets. In the second and third part, we provide an overview of the modeling framework used by WorldPop and the specific sampling approaches driving questions of interest. In the fourth part, we summarize the steps involved in the simulation experiment.

2.1. Study area and covariates

Because of the availability of fine spatial resolution census data, we selected Nepal as our study area. The population data were obtained from the Census Bureau of Nepal. Administratively Nepal is divided into 36,042 wards, 3647 local authority units, and 75 districts that are divided into five states. Based on natural conditions it can be divided into three belts: Terai (down belt), Hill (middle belt), and the Mountain Regions. These different belts also represent, respectively, the decreasing order of population density, as shown in Fig. 2A.

The enumerated residential population of Nepal as per 2011 census is 26,242,867. Based on administrative level 5 census data, the population density varies from 0 people per hectare to 1476 people per hectare, with mean and median density 9.04 and 2.78 people per hectare, respectively, indicating high skewness in the distribution.

The covariate datasets were prepared using ArcGIS (ESRI, 2016) and the Python programming language (version 2.7) (Python Software Foundation, 2013). The land cover is based on GlobCover data, which is derived from the ENVISAT satellite mission's MERIS (Medium Resolution Image Spectrometer) imagery. Land cover data were complemented by digital elevation data and derived slope estimates from SRTM-based HydroSheds data (Lehner, Verdin, & Jarvis, 2013). The Global Urban Footprint (GUF) 2016 data at 12 m resolution was collected from the DLR Earth Observation Center, and the Global Human Settlement Layer (Pesaresi et al., 2013) with a spatial resolution of 38 m was collected from the European Commission Joint Research Centre (2014 beta version). Observed lights at night as Visible Infrared Imaging Radiometer Suite (VIIRS) data (Hillger et al., 2013), within-country climatic spatial variations through the use of WorldClim/BioClim 1950–2000 mean annual precipitation (BIO12) and mean annual temperature (BIO1) estimates were also acquired (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005). In addition to land cover, settlement, and associated raster datasets, we included geospatial data that was correlated with human population presence on the landscape, such as networks of roads and waterways; large water bodies; settlement or populated locations; protected area delineations; and health facility locations (DSD Nepal, D. S. D. of N, 2015). All datasets were resampled using nearest neighbor to match same resolution to a square pixel resolution of 8.33 × 10–4 degrees (approximately 100 m at the equator) and projected into UTM 44.5 projection prior to analysis. Covariate data employed in the modeling process are summarized in Table 1.

Table 1
Data sources and variable names employed for population density estimation used for dasymetric weights.

Variable name(s)[a]	Description	Source and nominal resolution
Census	Nepal Census Data, 2011	Central Bureau of Statistics of Nepal, Admin-level 5
Land cover		GlobCover, 300 m (Arino et al., 2012)
cls011, dte011	Cultivated terrestrial lands	
cls040, dte040	Woody/Trees	
cls130, dte130	Shrubs	
cls140, dte140	Herbaceous	
cls160, dte160	Aquatic vegetation	
cls190, dte190	Urban area	
cls200, dte200	Bare areas	
cls210, dte210	Water bodies	
clsBLT, dteBLT	Built	
guf	Global Urban Footprint	DLR Earth Observation Center, 12 m
ghs	Global Human Settlement Layer	ECJRC, 38 m (Pesaresi et al., 2013)
lig	Lights at night	Suomi VIIRS-Derived (Hillger et al., 2013)
tem	Mean temperature, 1950–2000	WorldClim/BioClim (BIO1) (Hijmans et al., 2005)
Pre	Mean precipitation, 1950–2000	WorldClim/BioClim (BIO12)(Hijmans et al., 2005)
Pro	sanctuaries, national parks, game reserves, World Heritage Sites	World Database on Protected Areas September 2012, UNEP (UNEP-WCMC, 2010)
ele	Elevation	USGS HydroSHEDS (Lehner et al., 2013)
ele_slope	Derived Slope	USGS HydroSHEDS (Lehner et al., 2013)
hea_dist, hea_cls	Health Infrastructure of Nepal	Data Survey Department of Nepal
bui_dst, bui_cls	Building footprints	Open Street Map, 2017–07
res	Distance to residential areas	Open Street Map, 2017–07
pla_dst	Distance to places	Open Street Map, 2017–07
roa	Distance to road networks	Open Street Map, 2017–07
wat	Distance to waterbodies	Open Street Map, 2017–07

[a] The variable names are used in Random Forest model output and throughout the text as referenced to the specific data that they were derived from. The first three letters are derived from the data type (e.g. "lan" indicates land cover), and the last three letters, if present, indicate what type of data each variable represents (e.g. "_cls" is a binary classification, "_dst" is a calculated Euclidean distance-to, and "_dte" is Euclidean distance-to-outer-edge variable where positive distances are outsides and negative distances are inside areas).

2.2. Modeling framework

Fig. 1 provides a schematic diagram of the workflow used by WorldPop, based on the model used by Stevens et al. (2015). There are three submodel components: (1) Covariate preparation, (2) modeling and prediction of population density at the census unit level, which provides the weighting factor for redistribution in each census unit and (3) validation assessment. We used an RF (Breiman, 2001) to generate the predicted population density layer that informed our dasymetric redistribution of population counts (Stevens et al., 2015). RFs are a non-parametric ensemble modeling technique that uses bagging and a random selection of covariates across numerous classification and regression trees (Breiman, 1996) to reconstruct nonlinear relationships and interactions of the covariates (Breiman, 2001). When all trees are combined, the RF is robust to small and large sample sizes and "noisy" datasets (Breiman, 2001). Given that each tree is modeled independently, and is therefore parallelizable, modeling is efficient and there are minimal parameters to be set allowing for automation of the process (Liaw & Wiener, 2002). Bagging each new tree is fitted with a bootstrap sample of the training observations. The out-of-bag (OOB) error is the average error calculated using predictions from the trees and the remaining sample. This allows error to be computed for each tree while training the model.

A variance-stabilizing monotonic transformation is used to reduce the skewness and spatial heterogeneity of data. Stevens et al. (2015) found that by transforming the response variable using natural log prior to RF model fitting, consistently achieved higher prediction accuracies in the validation of aggregated predictions. Consequently, all the source datasets with zero counts were removed from the modeling process. A monotonic transformation of response variable does not alter the splitting rules for covariates. However, it can have a significant impact on model fitting and performance as it affects the values calculated for the splitting criteria objective function (sum of squared deviations for predictions in the RF regression case).

Using the model trained on the source scale, and predicting to the target scale using gridded covariates, a population density surface is predicted. The RF predicted population densities are used as relative weights to employ a dasymetric scheme (Mennis, 2003) to redistribute population counts within each source unit to the target cells (Stevens et al., 2015). The model used in this study differs from the model employed by WorldPop in dasymetric mapping part. While WorldPop model dasymetrically redistributes the population using the finest census level boundary (adm. 5), we dasymetrically redistribute the population using the next coarser census level boundary (adm. 4) and use the finest census level counts for validation. The weighted surface is summed up adm. level 4 (i.e., adm. 4) counts before dasymetrically redistributing the census counts into the ∼100 m grid cells.

2.3. Approach to generate and test spatial variation in sampling

Our motivation behind examining sample design was to better understand the possible mismatch that might existbetween training and prediction data under a change of scale. When the actual distribution of the response variable is unknown, the sampled data may vary from the actual distribution. In the case of aggregrated data, the mean is known but the variance of distribution will always vary across the scale. In such cases, there is a wider range of sampling within the known range and mean that could represent the possible scenarios. Applied to population data, the difference that exists between population density of urban and rural area means, it could follow bimodal distribution. Testing the spatial variation in sampling will also provide some insight on the behaviour of Random Forest model.

To design, such sample, the local spatial autocorrelation properties have been exploited. To determine the spatial relationship of each observation with its neighbor, the Moran's I of samples was calculated using a row-standardized weight matrix using three nearest neighbors. A Moran's plot represents the spatial relationship of each observation with its neighbor. Fig. 2C shows the Moran's plot of the transformed population density. Before the transformation, the global Moran's I was 0.41. Following log transformation, the global Moran's I increased to 0.65. In addition, the properties of a local indicator of spatial association (LISA (Anselin, 1995)) statistic, and corresponding standardized z-

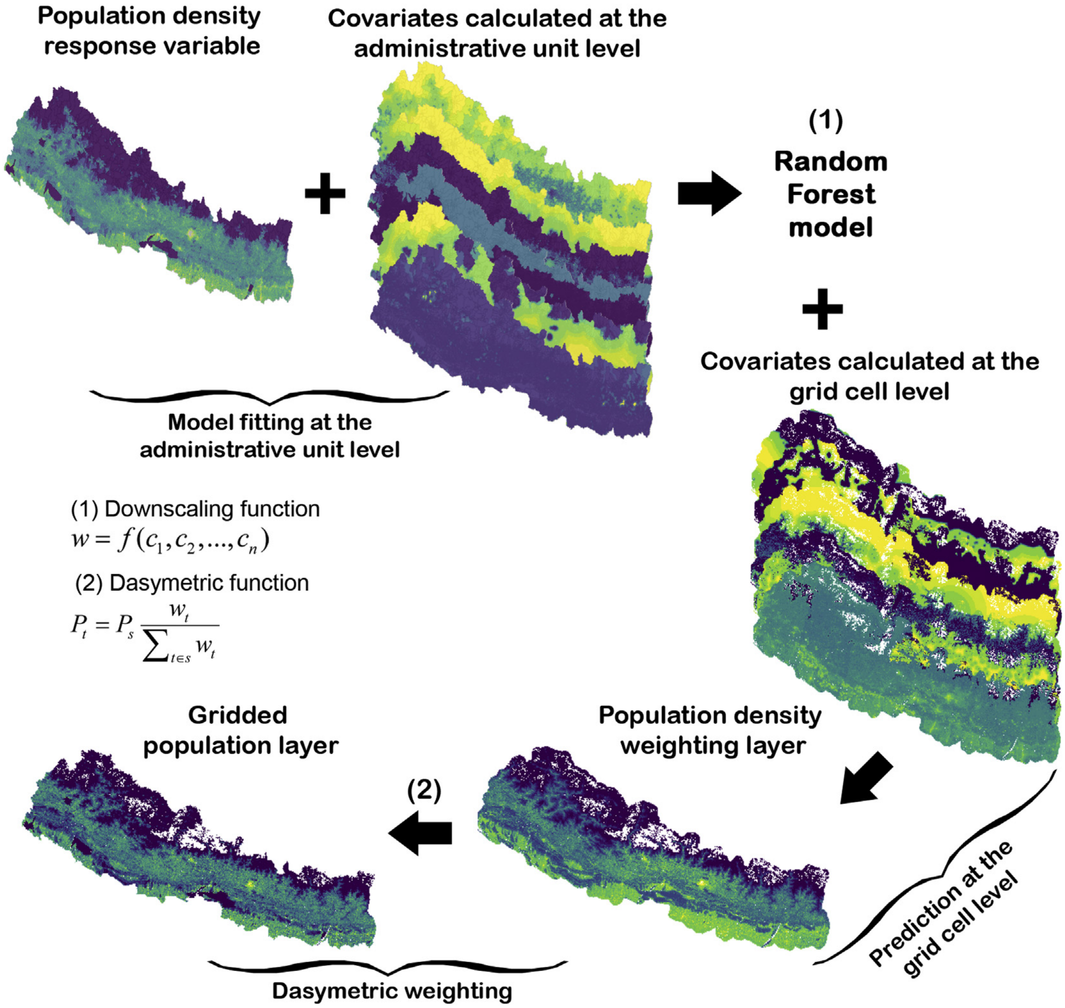
Population density response variable
Covariates calculated at the administrative unit level
(1)
Random Forest model
Covariates calculated at the grid cell level
Model fitting at the administrative unit level
(1) Downscaling function
$w = f(c_1, c_2, \ldots, c_n)$
(2) Dasymetric function
$P_t = P_s \frac{w_t}{\sum_{t \in s} w_t}$
Gridded population layer
(2)
Population density weighting layer
Prediction at the grid cell level
Dasymetric weighting

Fig. 1. Schematic diagram of the modeling process used by WorldPop. The covariates are prepared at the source and target scale. The source scale covariates are used for training whereas the target scale covariates are used for prediction of weighting layers for each grid. Using the dasymetric weighting function the final gridded population dataset is modeled.

scores and *p*-values provided increased nuance into the local level of variation in autocorrelation and heterogeneity of the different sample datasets.

Higher deviations in LISA statistics indicate the presence of spatial heterogeneity. Similar to Getis-Ord Gi* statistics based hot-spots and cold-spots (Getis & Ord, 1992), the clustering of high values and low values indicate pockets of non-stationarity with the origins present at the locations with the highest significance value. Shown in Fig. 2B, most of the clustering of low densities in the northern areas represent the harsh living conditions in the mountain areas, whereas the clustering of high densities in the capital region and southern plains represents the better living conditions and urban agglomeration. This is also reflected in Fig. 2C where points in quadrants I and III represent the clustering of high values with high-value neighbors and low values with low-value neighbors, whereas the points in II and IV represent the clustering of low values with high-value neighbors, and high values with low values neighbors. The significant wards in quadrant 'I' of Moran's Plot show high-high concentrations, where tiny units have very high population densities. These are areas either located in Kathmandu valley or relatively large cities, such as Pokhara, and represent the urban agglomeration. Less habitable areas in quadrant 'III,' having higher slopes and elevations, contain clusters of low populated areas, which are located in the northern and central parts of Nepal. The significant high-low wards represent areas in the north, with relatively lower population densities. These settlement units are located at high altitudes and are surrounded by valleys or peaks with inhabitable slopes, which require special attention. The significant low-high wards exist primarily in the south and have relatively higher population densities.

With that understanding, we designed the sampling to represent various levels of autocorrelation. The effective sample size calculation under an autoregressive specification for normally-distributed data (Griffith, 2005) suggested that given a Moran's *I* of 0.65 and a sample of 35,989 observations, about 14% of *iid* samples will have the same information. However, with a change in model specification, this value may differ to some extent. As the RF regression is non-parametric, this value could not be applied. Accordingly, we doubled the sample size to account for this, resulting in a fixed sample of 28% of all values, corresponding to 10,080 ± 10 units. The term *holdout* refers to unsampled units that were retained for validation.

The 28% sample set from the Moran's *I* calculation was generated using LISA statistics. We divided the census-based units into two mutually exclusive groups: significant and insignificant as determined by the LISA statistic at 0.1 level significance. A total of 11,127 (30.08%) were in the LISA significant group, and 24,862 units (69.92%) were in the insignificant group. We then adopted a stratified random sampling scheme, fixing the percentages of significant and insignificant units in the total sample to generate samples with a varying level of autocorrelation, as detailed in Table 2. For each level of autocorrelation, the sampling procedure was repeated 100 times, calculating and recording

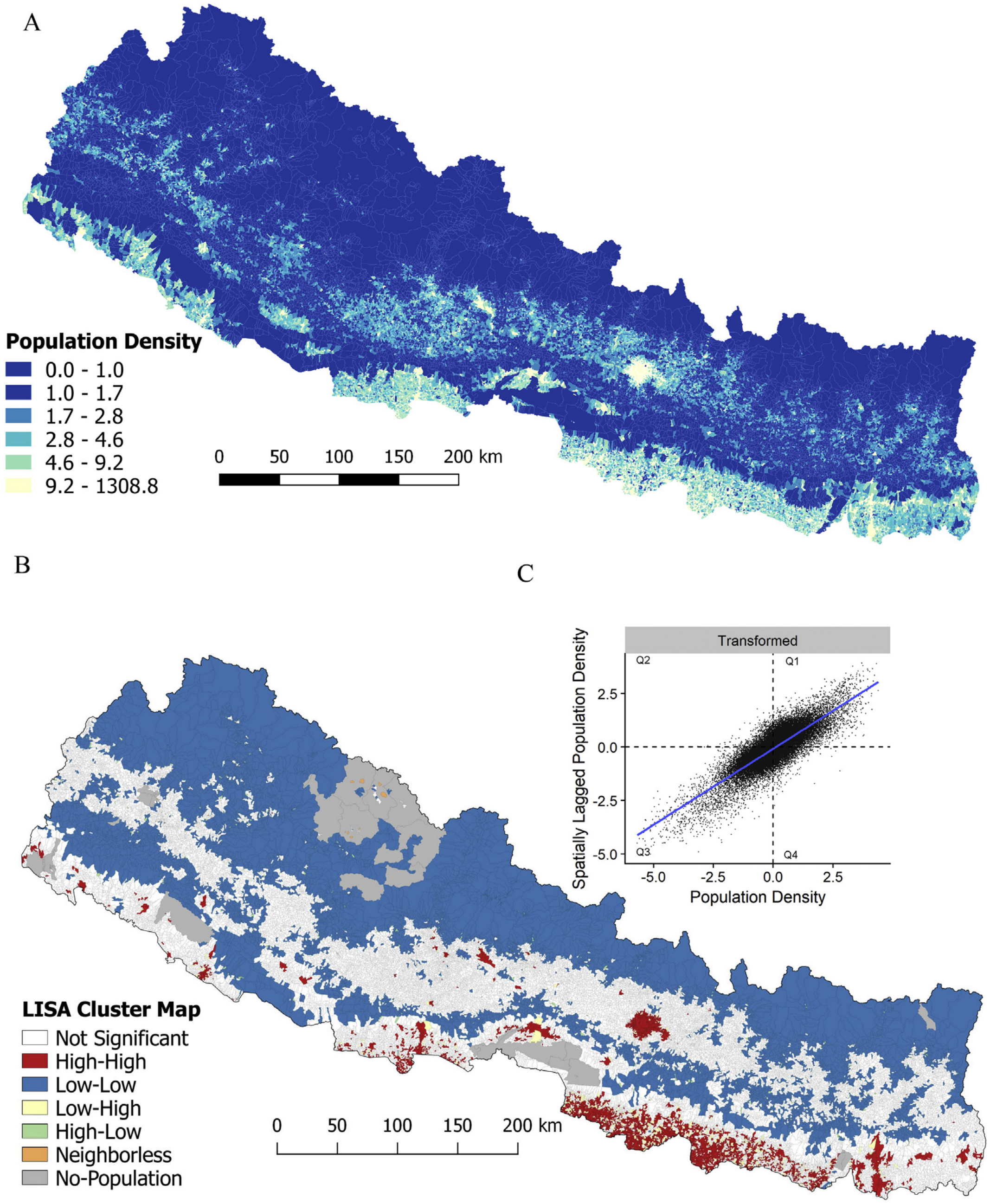

Fig. 2. (A) Average population per hectare in 36,042 administrative units of Nepal represented in six quantiles (B) statistically significant high-high, low-low, low-high, and high-low regions. Regions with no population, such as national parks, high peaks are excluded (C) Moran's plot of transformed population density. The extreme values existing in the non-transformed data became more homogeneous following log transformation. The x-axis represents the standardized response variable, and the y-axis represents the average of the standardized neighboring values.

Table 2
Moran's *I* of the base dataset and average Moran's *I* across samples. % significant LISA units column represents the percentage of LISA significant units in the sample.

Model	% Sig. LISA units	Moran's *I*	Model	% Sig. LISA units	Moran's *I*	Model	% Sig. LISA units	Moran's *I*
Base	–	0.65	Sample 5	25	0.61	Sample 12	60	0.79
SRS[a]	–	0.65	Sample 6	30	0.65	Sample 13	65	0.81
GSRS[b]	–	0.70	Sample 7	35	0.68	Sample 14	70	0.82
Sample 1	5	0.37	Sample 8	40	0.71	Sample 15	75	0.83
Sample 2	10	0.45	Sample 9	45	0.73	Sample 16	80	0.85
Sample 3	15	0.51	Sample 10	50	0.75	Sample 17	85	0.86
Sample 4	20	0.57	Sample 11	55	0.77	Sample 18	90	0.87

[a] Simple random sampling.
[b] Geographically stratified random sampling.

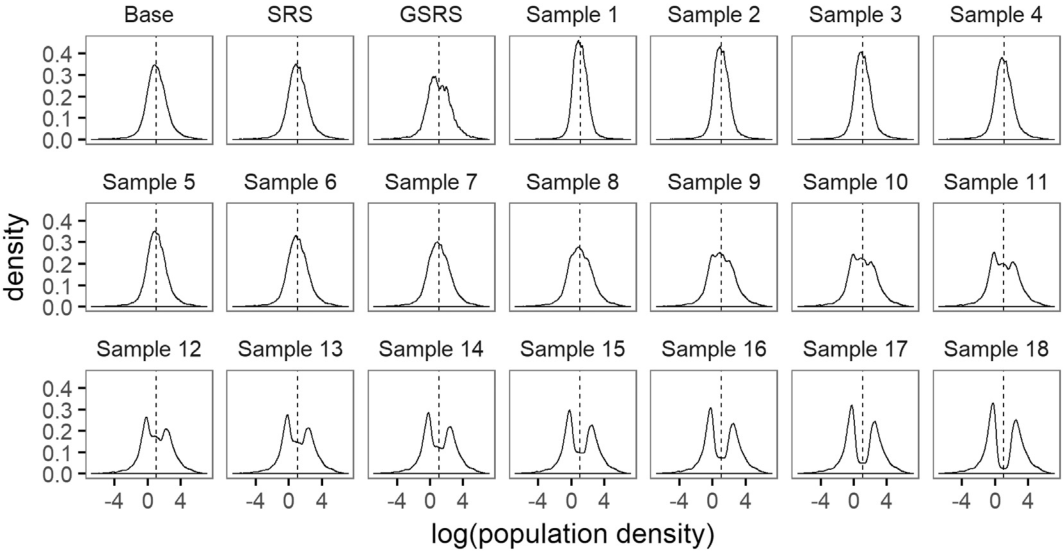

Fig. 3. Density plot of log-transformed distributions for the base dataset, simple random sample(SRS), geographically stratified random sample (GSRS), and eighteen samples at different Moran's I.

Moran's *I* for each repetition, and calculating the corresponding average Moran's *I*, across all repetitions, for each level (Table 2). The average Moran's *I* across the 100 samples of each variation is shown in Table 2.

2.4. Simulation and validation

A general overview of simulation framework is shown in Fig. 4. The steps involved in generating the samples and random forest prediction and validation is summarized in the flow diagram. As portrays in the process chart of a simulation experiment, the dataset is divided into two groups based on LISA p-value, and 10,080 samples were drawn from each of these two groups by varying the ratio from 5% to 90%. For SRS, 10,080 samples are randomly selected. The GSRS is based on 10,080 equal size hexagons, and one administrative region is randomly selected inside each hexagon. For LISA based samples, the administrative units not selected in the sampling process constitutes a holdout dataset that is used for validation. The RF model is trained based on a given sample dataset using the census level covariates. The corresponding holdout dataset for each LISA sample is used to validate the trained model.

Using the sample-based RF model and grid level covariates, we predict the population density at the grid scale. Lacking a finer-scale dataset to validate grid scale predictions we compared these predictions to the base model (trained on all census-based units). The predicted grid-based population density data is used as a weighting layer for dasymetric redistribution. The predicted data is scaled to match the population counts at the level 4 data, i.e., the next coarser census-based population count. Summing the grid-based population counts within each of the finer-scale census-based units level 5, we then compared the aggregated counts to the original census counts to approximately ascertain the RF's capture of fine-scale population distributions. For comparision, we used Root mean square deviation (RMSD), %RMSD, mean absolute deviation (MAD) and the r-square (RSQ). The RMSD is the square root of the mean squared deviation between observed and predicted. The %RMSD is the square root of the mean squared deviation of residuals between observed and predicted divided by the observed value. The MAD is the mean of the absolute deviation of residuals between the predicted and observed values. The r-square is a measure of goodness of fit that indicates the linear association between the predicted and observed values.

3. Results

The results are based on 2000 samples representing 100 replications of 18 sample types with varying levels of Moran's *I* alongside 100 samples each from the simple random samples, and geographically stratified random samples. We discuss the properties of these different sampling approaches, the variations in the model training, and the prediction results. Important to note is that the Moran's *I* of the source data (i.e., administrative unit, n = 10,080), 0.65, is considered to be the base level of autocorrelation. As such, Moran's *I* higher than 0.65 indicates high autocorrelation and lower than 0.65 indicates low autocorrelation. To visualize the effect of sample types with higher and lower levels of Moran's *I* on RF predictions in detail, out of eighteen we selected three samples that represented lower (sample 2), similar (sample 6), and higher-level (sample 16) Moran's *I* compared to the base dataset Moran's *I*. For consistency across the results we have used the log-transformed values of the population densities.

The simple random sample was generated by randomly selecting the samples, while the geographically stratified sample was based on binning the area using hexagons, in which one unit was selected from each hexagon. The average Moran's *I* value of simple random sample (0.65) was like the base level but the average Moran's *I* for the geographically stratified sample was 0.68. For the additional eighteen sample types, there is a varying composition of significant and non-significant LISA units that controlled the autocorrelation captured by Moran's *I*.

Sample 6 (n = 10,080) and base dataset (n = 35,989) both have 30% significant LISA units (Table 2), and sample 6 has a composition similar to samples that were obtained using simple random sampling. The average Moran's *I* of sample 6 and SRS were same as the Moran's *I* of base dataset. With a decrease in units with significant LISA, the average global Moran's *I* decreased, and the peak of the distribution was pushed upwards with a lighter tail. On the other hand, an increase in units from the group with significant LISA, the average global Moran's *I* increased, and the peak of the distribution was pushed downwards with a heavier tail. *I*nterestingly, when 90% of the sample was composed of significant units, the distribution became bimodal, creating a valley centered on the mean. The two modes of this distribution represented the mean of “high-high,” and the “low-low” clustered LISA values (Fig. 2B). The distribution of samples across different levels of autocorrelation is depicted in Fig. 3.

The variation of the sample means and its variance with variation in the level of the autocorrelation of samples is shown in Fig. 5A and B.

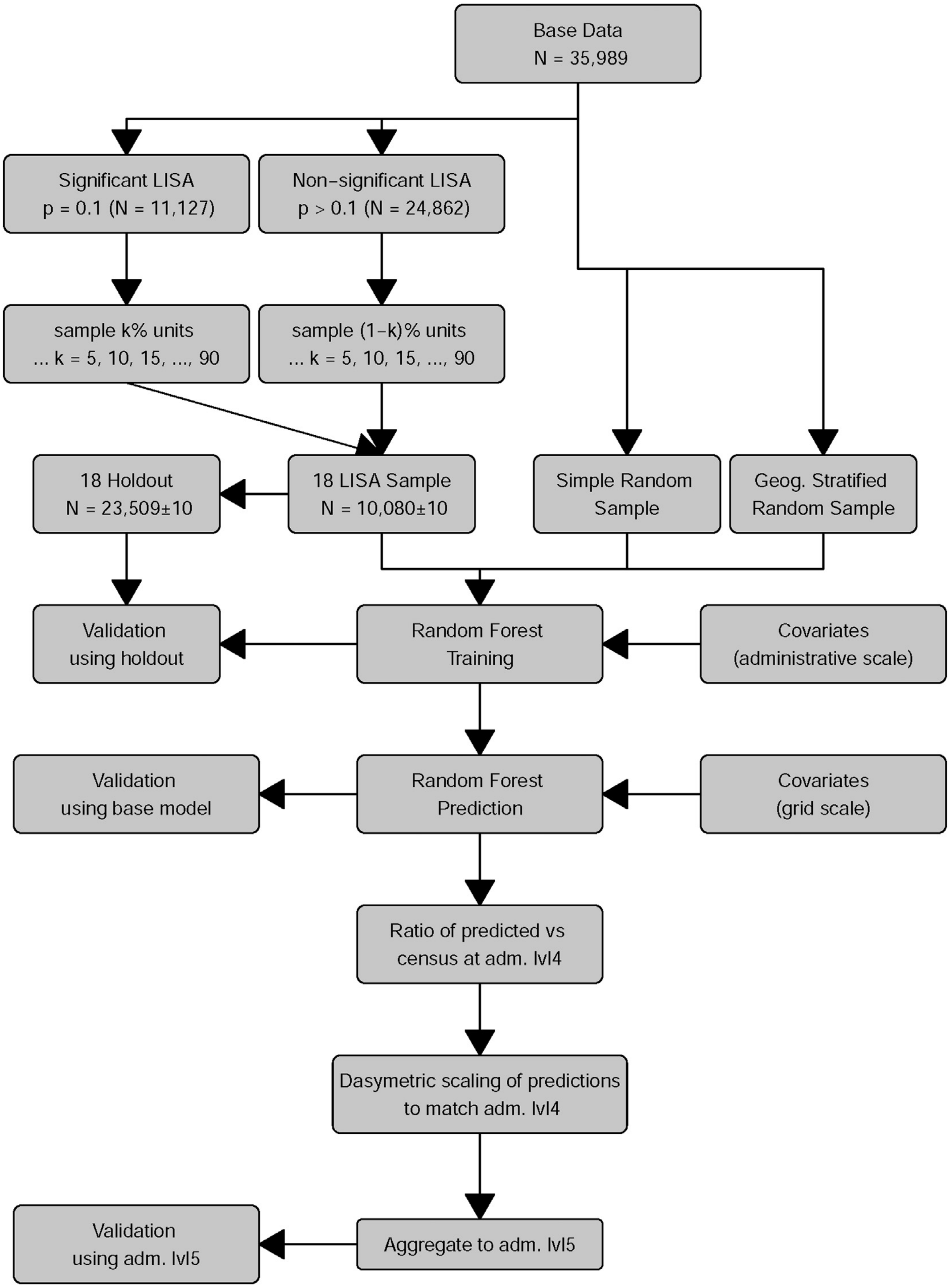

Fig. 4. Process chart of the simulation experiment.

The grey dots, red dots, and blue dots represent eighteen hundred LISA-based samples, one hundred each of simple random samples, and geographically stratified random samples respectively. Change of global Moran's *I* in samples affected the variance more than the mean. Samples with higher Moran's *I* had higher variance than lower Moran's *I* samples. This observation was consistent with the effect of autocorrelation on sample variance (Griffith, 1992, 2005). The LISA based samples and random samples both had a similar sample mean-variance with mean 1.062. However, the geographically stratified random sampling was highly biased, with mean 0.36.

The variance explained by the models based upon the OOB[1] samples showed a curvilinear increase (Fig. 5D) with an increase in Moran's *I*; however, the mean square error (MSE) begins to decrease when the non-significant component in the sample was 20% or less (Fig. 5C). While validating using holdout units the RMSD values showed an opposite pattern to the OOB units, but for higher ranges of Moran's *I*, R-square between holdout and predictions decreased rapidly. The Moran's *I* of residuals indicate a spatially structured relationship across the

[1] The OOB error estimates are generated internally by the RF model. Each tree in randomForest (Liaw & Wiener, 2015) R packages is constructed on 63% bootstrap training data, leaving 37% training data for validation.

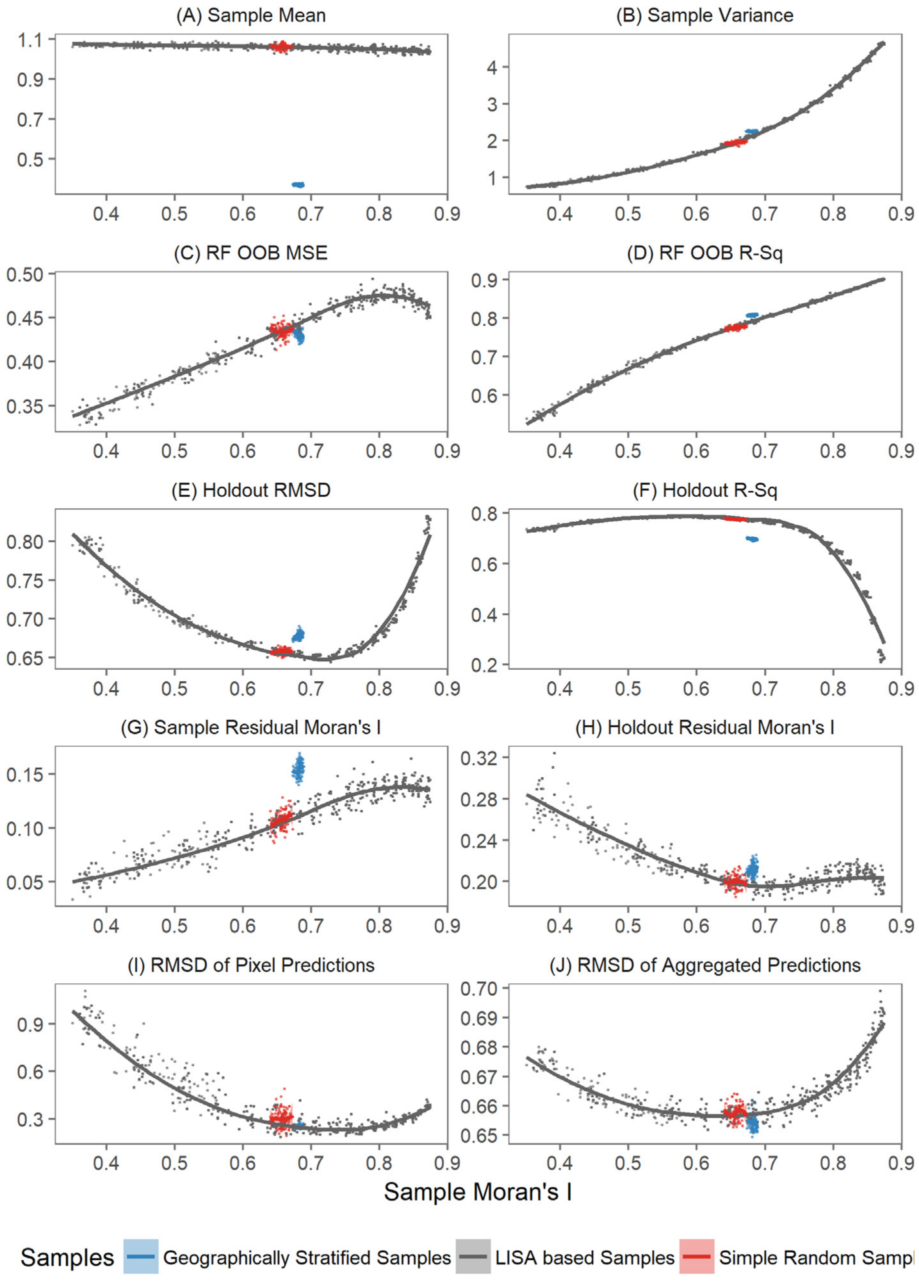

Fig. 5. All samples are plotted in each panel representing the geographically stratified samples (blue) LISA based samples (grey) and simple random samples (red). The X-axis represents sample Moran's I and Y-axis represents: (A) mean of sampled units for training at census level, (B) sample variance of sampled units for training at census level, (C) training Model MSE based on OOB, (D) training model R-square based on OOB, (E) RMSD of holdout (F) R-square of holdout (G) spatial autocorrelation in the residuals of training samples, (H) spatial autocorrelation in the residuals of holdout, (I) Root mean square deviation of predictions from a full set of census data, (J) Validation of transformed population density following dasymetric weighting at census scale with actual population density. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

study area. The residuals of the training model and validation with holdout units, in all cases (Fig. 5G and H), exhibited positive autocorrelation. The residuals of holdout units showed a decreased autocorrelation with a minimum Moran's *I* of approximately 0.7. However, this value was always higher than the training residuals. The RMSD values of holdout units across all the values of Moran's *I* had a ‘V’ shape. This metric was lowest near the 0.7 Moran's *I* level, which was slightly higher than the base level.

While validating using holdout units, the R-square and RMSD values for lower ranges of the sample Moran's *I* showed a similar pattern to the OOB units, but for higher ranges of Moran's *I*, the predictions showed lower errors. R-square was also robust for higher ranges of Moran's *I*. This was contrary to the observations with the holdout data and indicated that RF was producing better predictions with higher autocorrelation in the training data (larger Moran's *I*). Models estimated with samples having lower Moran's *I* had a smaller range for predictionand those with higher Moran's *I* had a more extensive range. We compared the population datasets produced by dasymetrically redistributing administrative level 4 census counts with the counts of the finer, administrative unit level 5 data. (Fig. 5I). The dasymetric process was sensitive to the spatial heterogeneity of population counts that was present at the administrative unit level (Fig. 5J).

3.1. Detailed investigation of training and prediction

The three selected samples from each LISA based sampling: sample 2, sample 6, and sample 16 are referred as MI-.45, MI-.65, MI-.85, representing lower, similar, and a higher level of Moran's *I* than the base level. In Fig. 6 these three samples are shown in red, while the corresponding holdout is shown in grey. As the autocorrelation was increased, the concentration of samples began to move from the center of the distribution towards the tails. With the highest level of autocorrelation, again, displaying a bimodal distribution (Fig. 3). The holdout units showed different patterns. As the peak of the sampling distribution began to decrease with higher autocorrelation, the peak of holdout units began to increase, showing that the units had a low

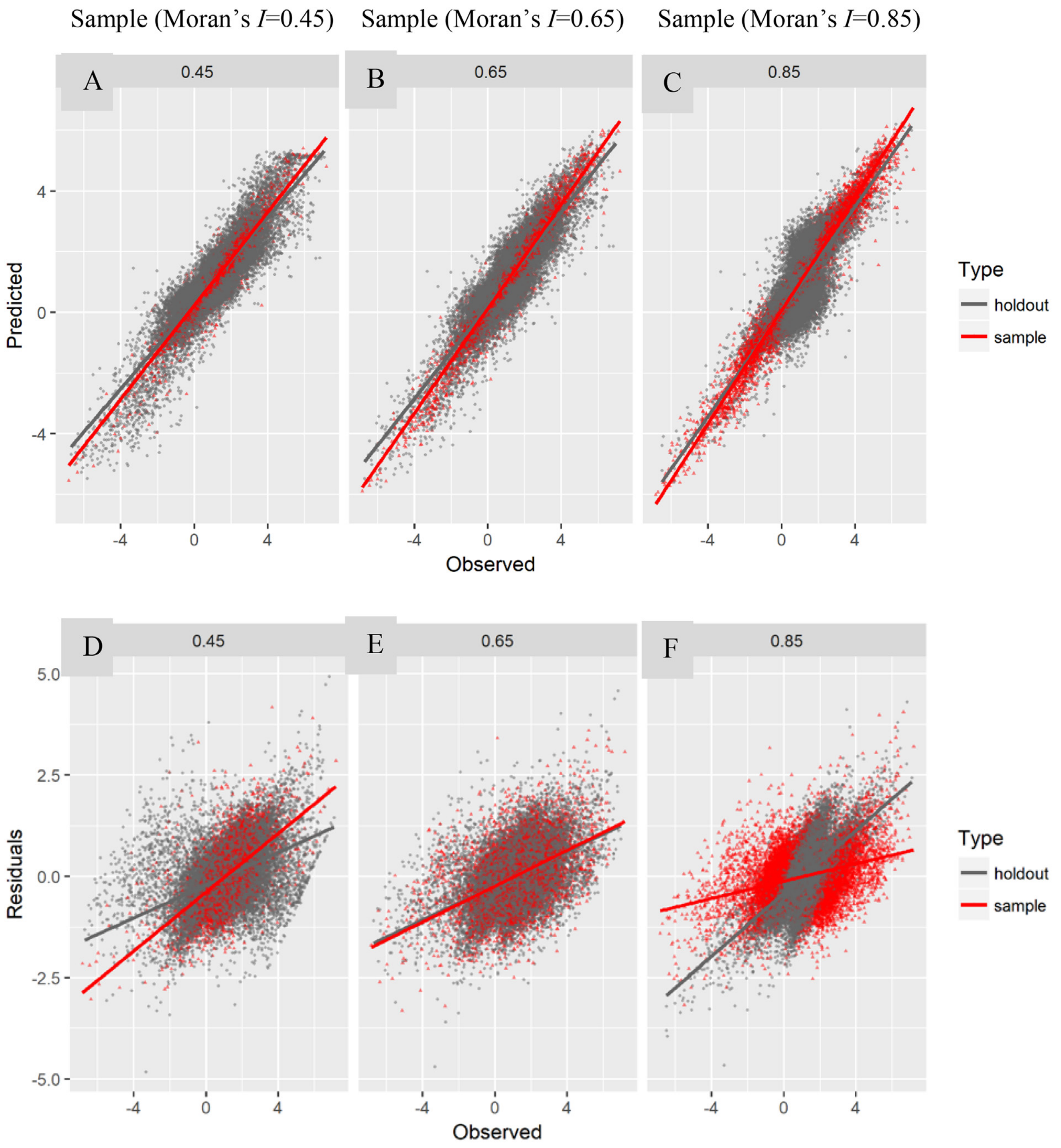

Fig. 6. A, B, and C show predictions by samples with various levels of autocorrelation. D, E, and F show residuals by samples with various levels of autocorrelation. For Morans I 0.45sample, a significant part of the sample consisted of values that surrounded the mean, whereas the maximum part of holdout have extreme values. For Moran's I 0.65 both training sample and holdout have similar distribution of extreme values. The Moran's I 0.85 sample has opposite distribution to MI0.45. The output of model A trained on the lower autocorrelation sample tended to predict within a narrow range around the mean and the holdout residuals has high range. The output of B and residual have similar variability. The output of C have higher variability and residuals have a less variability. The model trained on this type of sample overpredicted the values that were higher than the mean and underpredicted the values that were lower than the mean.

deviation from the mean.

3.1.1. RF model

The effect of the RF model was captured by the plots shown in Fig. 6A, B, and C. Fig. 6A shows the observed and predicted values of the sampling units and holdout units under low autocorrelation. The predicted values for high and low values deviated in higher magnitude from the fitted line. Fig. 6B showed the case with a similar level of autocorrelation as the base dataset and had a lower bias. Fig. 6C shows relatively higher autocorrelation samples, where the middle values showed values nearby the mean, having an 'S'-like shape. The slope of the fitted line increased with higher autocorrelation, which showed the better goodness of fit, and the difference between the slope of holdout units and sample units became smaller with an increase in the sample autocorrelation. Across these four models, the mean of squared residuals is the lowest for the base model (Table 3), but R square is highest for the sample with Moran's *I* 0.85.

A negative residual represents an over prediction, whereas a positive residual represents underprediction. Ideally, the residuals should have a random pattern but, these revealed a linear relationship between the observed values (Fig. 6D, E, F) indicating residuals are not white noise and still contain information. As the units not selected as part of a

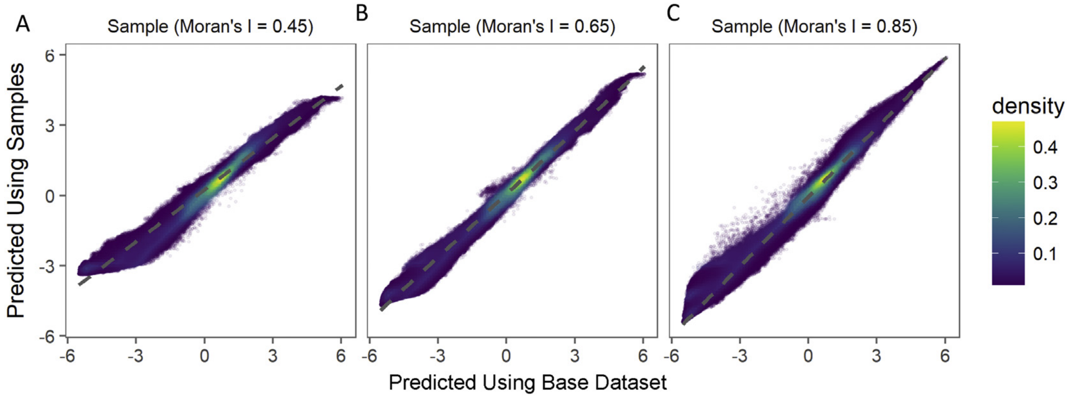

Fig. 7. Comparison of predictions from three selected samples and predictions from the base dataset (trained using all the census units). The top panel indicates the Moran's I value of training data. For Moran's I 0.45 the predicted values at pixel scale are under predicted for high and low values. For 0.65, the predicted values are still under predicted for some of the extremely high and low values. For 0.85, overall prediction is spread across the range. Relative goodness of fit for 0.85 is therefore the best relative to the model fit with the complete, base dataset.

sample are used for holdout units these necessarily share an inverse relationship in their Moran's *I*. If the Moran's *I* of the sample is lower the holdout Moran's *I* is higher and vice-versa except for samples with Moran's *I* matched to the base dataset. The slope of residuals of sample units decreased with higher sample autocorrelation while the slope of residuals of holdout dataset increased. A mismatch in training dataset Moran's *I* resulted in a difference in the unexplained part of the residuals.

Higher RF variable importance scores for any covariate indicates a decline in mean squared error (MSE) of prediction using the observed covariate values compared with predictions when that covariate's values are randomly shuffled. It is important to note that a variable importance score may change with each independent simulation if collinearity exists among the predictors. The overall percent explained MSE of an RF model reduces with the decrease in samples size. While distance from residential areas is the top variable in the base model, it is the top variable for samples with low and similar Moran's *I* but third important variable for higher autocorrelation sample. Distance from roads is the top predictor for higher autocorrelation samples while it is the second most important predictor for base sample and MI-.65 samples. Distance from the edge of GUF built-area is also a top predictor except for the MI-.45 samples. Elevation and slope are among the top five important predictors for base dataset but not for samples.

3.1.2. Grid-scale prediction

Here, we used the base model as a benchmark to compare the sampling performance of the three selected samples. Fig. 7 shows the predictions based on MI-.45, MI-.65, MI-.85 models as described in Fig. 8 compared with predictions using all the census units.

Table 3
Validation of RF model at source scale.

Population counts	Base	MI-0.85	MI-0.65	MI-0.45
Mean of squared residuals	0.37	0.35	0.41	0.44
Percent variance explained	80.75	62.68	78.21	89.57

As the sample autocorrelation is increased, the regression line of model prediction gets aligned with the prediction of the base model in Fig. 7. Also, we find that more values are above the regression line, meaning the values are being overpredicted compared to the prediction using the base model. Higher autocorrelation in the sample data increases the extreme values in the sample. At autocorrelation level 0.85, the bimodal distribution of sample has higher variance. Random forest predictions have a higher range at this level. The predictions from higher autocorrelation samples have better predictions than two other lower autocorrelation samples. The range of predictions of these three samples varies up to 4.1, 5.5, and 6.1 in log scale for Moran's *I* 0.45, 0.65, and 0.85. The higher range of prediction at pixel scale ensures better dasymetric redistribution of population and better estimates the extremes of the distribution of the observed, base dataset.

The validation assessment based a population map informed by coarser census-level data (adm 4) compared back to finer census scale counts (adm.5) is shown in Table 4. In terms of higher accuracy, the model with higher Moran's *I* lags behind the base model. As noted by

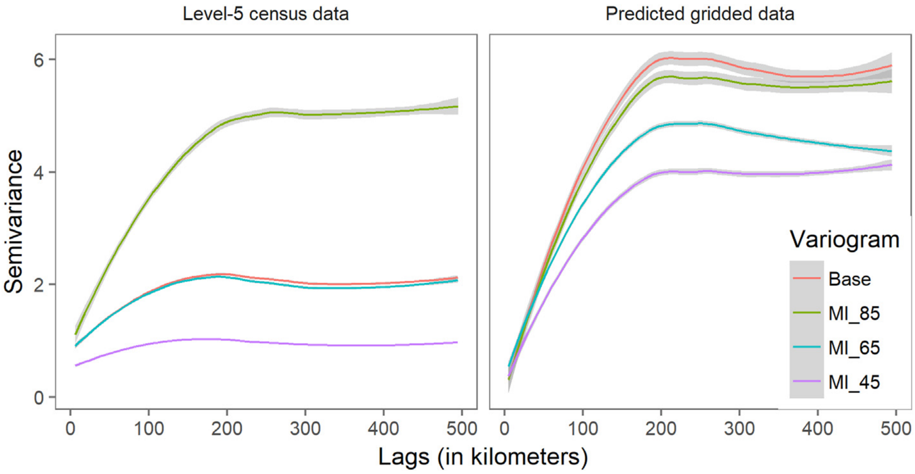

Fig. 8. Comparison of semivariance of samples and predicted data calculated at 10 km binned intervals. The MI here indicates Moran's I and the two digits used indicates the value of Moran's I of the samples for e.g. MI-45 indicates sample with Moran's I value 0.45. The sill of the sample increases with the increase in the autocorrelation. The MI-.65 has similar sill and range compared to the base data. The gridded predictions from the base model have the highest sill.

Table 4
Validation of predictions counts at target scale.

Population counts	Base	MI-0.85	MI-0.65	MI-0.45
"MAD"	285.08	298.24	291.33	294.73
"%RMSD"	85.14	89.41	88.84	94.47

Stevens et al. (2015), the Random Forest model based dasymetric approach tends to predict very small values in the rural and less dense areas. Using a fractional continuous surface population counts for rural areas leads to underprediction of population density and higher amount of errors.

Compared to the simple random samples, the predictions from geographically stratified sample had higher RMSE and lower R-Square value both at the source scale and target scale. However, the RMSD values of simple random samples are comparatively scattered. The geographically stratified sample have higher autocorrelation than simple random samples, but due to the underrepresentation of extreme values, the heterogeneity is reduced. The trained model on this sample underpredicts the high population densities areas both at the source and target scales. Due to the stratification, each replication of geographically stratified samples is evenly distributed across the country and captures a similar level of variability while the variability of simple random sample varied with each replication. This variability of the simple random sample is captured by the scattering of the RMSD.

4. Discussion

In this study, we analyzed the error propagation caused due to the mismatch from the source scale to the target scale using an RF-based dasymetric model to produce a gridded population dataset. The eighteen different types of samples with varying levels of spatial autocorrelation represented the possible mismatches in source and target scales. These were referenced against typical sampling techniques of simple random sampling and geographically stratified sampling. The findings of this experiment can be summarized into the findings at the census unit scale (using holdout units) and findings at grid scale.

4.1. Impact on RF training and prediction at source scale

At the source scale, the samples with lower Moran's *I* had a higher proportion of non-significant LISA units. All non-significant units are small towns or villages with varying population density levels. As shown in Fig. 7A, the model trained on this data tends to deliver poor predictions, for the observed, high-density and very low-density areas that are tightly clustered about the mean. Samples that have a higher Moran's *I* than the original census data, and a higher proportion of significant LISA units, tended to originate from areas containing large cities (Fig. 2C, quadrant I and III) with high population density or from areas of very low population density settlements in the mountains or isolated in plains. Further, models trained on this data tend to over predict or under predict the values in the middle range. With higher autocorrelation, the OOB samples resulted in the higher goodness of fit. However, the MSE value started falling at extremely high Moran's *I* values which is likely due to the reduction in the number of units with high population density in the sample and MSE being sensitive to outlier errors. The non-random pattern of training model residuals had a linear relationship with the observations, and the spatial pattern of residuals became more evident with higher levels of sample Moran's *I*. The residuals of the RF model, trained on a full set of observations also had spatially autocorrelated residuals, which is not surprising since RFs, given their ensemble nature, cannot predict outside the observed range of the response variable (i.e., they always regress towards the mean to some degree (Breiman, 2001)). If high population density units in the original census data are clustered in space, then under prediction in those clustered areas (over-prediction in less densely populated units) will result in significant residual spatial autocorrelation. The varying coarseness of census units added uncertainty and increased errors in the RF model predictions. Also, the higher the autocorrelation, the more similar the fit produced from each bootstrap sample and result in a high variance of Random Forest model.

The simple random samples have mean, variance and Moran's *I* similar to the base level dataset. Hence it captured the exact level of heterogeneity and autocorrelation. The geographically stratified random sample that was based on regular hexagons, however, captured more variation in population density across census units since it tends to oversample the larger, sparsely densely populated units relative the higher number of smaller, high-density units present in the original data. Hence the bias of the geographically stratified sample reflected the uneven spatial distribution of the population. At the target scale, the prediction from the simple random sample had higher RMSE compared to the geographically stratified sample. This phenomenon could be understood by observing the difference in the mean and range of training and prediction data. The mean of geographic random sample training data was much less than the simple random sample mean, hence the RMSE is concentrated showing the similarity of predictions.

4.2. Impact on RF prediction at target scale

This study also fits into the larger class of work on change of support problems. In geostatistics, the change of support problem addresses a large set of problems where the observations and estimation are done at different scales. Detecting the varying nature of a particular phenomenon depending on the geographic scale of analysis is well documented as far back as the early 1950s. Robinson (1950) pointed out that due to strong spatial effect, inference about individuals based on group-level data could be contradictory which is also known as an ecological fallacy. Under a change of support, the variance at source scale represents the variances of mean at the target scale. In geostatistics, this phenomenon is explained by the regularization theory (Journel & Huijbregts, 2003) that explains the change in the variogram of a spatial attribute as support changes. The general result is that the sill of the variogram decreases as one moves from finer to coarser support and that this effect is stronger when the nugget variance is significant. A drop of the sill means that the spatial variability decreases, and this agrees with the observation that coefficients of variability often decrease with higher levels of spatial aggregation (Atkinson & Tate, 2000; Dumanski, Pettapiece, & McGregor, 1998). In a spatial interpolation context, the averaging-out effect causes the coarser scale kriging variance to be smaller than the finer scale kriging variance.

Fig. 8 portrays the empirical semivariance of source scale and corresponding target scale prediction using the base dataset and samples with higher, similar, and lower autocorrelation. Among three samples, MI-.85 have a higher sill. As we find in Fig. 8, although the MI-.65 has similar sill and range, the prediction from MI-.85 is closer to the predictions from the base model. In a top-down population modeling, with higher autocorrelation in the training samples, the variance of samples at the source scale is much higher and possibly closer to the variance at the target scale. For target scale prediction, using a sample with lower Moran's *I*, high-density areas were under-predicted to the mean and low-density areas were overpredicted (Fig. 7A). Predictions from training data with a base level of Moran's *I* also have similar distribution (Fig. 7B). On the contrary, training data with higher Moran's *I*, the predictions had relatively balanced distribution as shown in Fig. 7C. This balance is due to the presence of both high and low values that resulted in higher variance. After the dasymetric weighting process, these predictions were aggregated and compared to the actual census counts. In the comparison (Fig. 7C) we find more accuracy for the higher ranges of autocorrelation but not for, the lower. However, the current method of validating the accuracy of the predicted output maps, based on aggregated counts, leaves a wider scope for errors in the disaggregated predictions.

Spatial downscaling that relies on the use of ancillary variables is a focus of research in many fields, but determining the accuracy of such downscaling is often limited due to the absence of validation data (Addiscott, 1998; Heuvelink, 1998). Few studies related to downscaling in spatial ecology and environmental modeling have analyzed the effect of scale across different nested spatial hierarchies (Addiscott, 1998; Chave, 2013; Gustafson, 1998; Heuvelink, 1998). Gardner, Mime, Turner, and Neill (1987), concluded that if the underlying distribution is totally random, data at different scales could be used as a neutral source for modeling, however, in the presence of a spatial structure in a dataset, the results might differ across those scales (O'Neill, Gardner, & Turner, 1992; Wu, Jelinski, Luck, & Tueller, 2000). Heuvelink (1998) divided the uncertainty associated with a top-down modeling approach into two components: input errors and modeling errors. The input error is support dependent, meaning that the error of the source average will be smaller than the error at any given target within the source.

The modeling error results from various assumptions, discretizations, and simplifications that are made to make the model manageable, which is often difficult to quantify. Heuvelink (1998) suggested validation is a reliable method to assess the *model error* contribution, which involves comparison of model predictions with independent measurements resulting in the output errors. When input and model errors are statistically independent, the variance of the model error could be estimated by subtracting the variance due to input errors from the variance of output errors. However, the main problem is directly obtaining the data against which to test the model and with the increase in the hierarchical level, the uncertainty attached to the distribution increases as well that makes model evaluation correspondingly more difficult and less satisfactory (Addiscott, 1998; King, Fox, Daroussin, Le Bissonnais, & Danneels, 1998). An alternative for model evaluation would be evaluating the model against aggregated data at the scale of the source and assuming that the evaluation holds good at target levels (Radcliffe, Gupte, & Box Jr., 1998; Stevens et al., 2015). Results of Radcliffe et al. (1998) suggest that this method could be reasonable, but the sum of residual will be smaller than that of the target scale.

5. Conclusion

This study contributes to the literature on the application of ensemble models under spatial autocorrelation and spatial downscaling that deals with interpolating aggregated counts to a finer scale. The errors in gridded data can be attributed to misspecification of spatial scale, and ecological fallacy as the relationships from inter-variability of source scale is used to model intra-variability inside each source at a target scale. Modeling finer scale distribution with courser scale training data means will lead to underprediction of extreme values (i.e. regression towards the mean). Due to the change of support, the range of population density of the aggregated training data and the zonal mean of covariates lead to a smaller range of training. The predictions could be misleading when predicting new observations with values that were not seen in the training set. The established relationship and range of variation might change with size and shape of areal units. Also, the limitation of a tree-based model could compound this issue, as the prediction from a tree-based model is limited to the range of the training data. All the predictions using the target scale covariates that have values outside the range of the training covariates are predicted upto the range of training scale. By taking avarage of these values, the range of target scale prediction from Random Forest model will always smaller than the range of training scale observations. The captured nonlinear relationship by the RF model will differ from a model trained on samples of disaggregated target level datasets. A third limitation of this study is related to the exclusion of the covariate errors such as measurement and positional error and impact of upsampling of covariates to higher resolution, e.g., ESA land use dataset was downsampled from 300 m resolution to 100 m.

The problems motivating this study are not unique to gridded population modeling or spatial downscaling but several disciplines that use RF modeling for spatial datasets. Our results suggest that for samples of spatially-explicit data used for prediction from the RF model at the same scale, the presence of spatial autocorrelation leads to high variance of the residuals. The RF model is sensitive to the mismatch of spatial autocorrelation of the training sample, and a representative sample of the actual population is helpful to achieve the best fitting models. In many real-world scenarios, when the spatial distribution of a variable is unknown, precaution has to be made in selecting the samples as the mismatch could potentially affect the training. Some research has been done in the area of spatial data mining aimed to be used with spatial data with high autocorrelation and heterogeneity such as spatial decision trees, and spatial ensemble learning (Jiang, Li, Shekhar, Rampi, & Knight, 2017; Jiang & Shekhar, 2017). These models use a spatial measure deduced from the spatial weight matrix as a variable in the modeling process; however, scalability remains an issue. Some other methods such as spatial resampling and block bootstrapping have been used to capture the spatial structure (Brenning, 2012; Goetz, Brenning, Petschko, & Leopold, 2015).

RF predictions from higher autocorrelation samples were found to be better distributed and spread out at target scale. However, the unavailability of validation data forbids us to make a generalized statement in this regards. This work can be extended in different ways. One extension could be improvising the model based on upper and lower bounds of variance and range information at the target level. A possible solution could be formulated by collecting target scale validation datasets and examining its correlation with auxiliary variables. Another possible extension could be the use of recent large-scale detection of building footprints information from satellite imagery (Tiecke et al., 2017; Yuan, 2016). Finally, the allocation process of the population could be improved by using discrete allocation techniques.

All computation in this article was produced using R statistical computing environment (version 3.4.1) (R Core Team, 2017) with R packages raster (Hijmans, 2016), sp. (Pebesma, Bivand, Rowlingson, & Gomez-Rubio, 2013), data.table (Dowle et al., 2018), random Forest (Liaw & Wiener, 2015), Rborist (Seligman, 2016), and ggplot2 (Wickham, 2016). Programs and data are available from the corresponding author on request.

Acknowledgments

This work is supported by the funding of the Bill & Melinda Gates Foundation (OPP1134076). A full set of data sources and acknowledgments are provided on the WorldPop website (www.worldpop.org). Parts of this work has been conducted using the resources of the University of Louisville‘s research computing group and the Cardinal Research Cluster.

References

Addiscott, T. M. (1998). Modelling concepts and their relation to the scale of the problem. *Nutrient Cycling in Agroecosystems, 50*, 239–245. https://doi.org/10.1023/A:1009796413132.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis, 27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Arino, O., Ramos Perez, J. J., Kalogirou, V., Bontemps, S., Defourny, P., & Van Bogaert, E. (2012). *Global land cover map for 2009 (GlobCover 2009)*. https://doi.org/10.1594/PANGAEA.787668.

Atkinson, P. M., & Tate, N. J. (2000). Spatial scale problems and geostatistical solutions: A review. *The Professional Geographer, 52*(4), 607–623. https://doi.org/10.1111/0033-0124.00250.

Azar, Derek, Engstrom, Ryan, Graesser, Jordan, & Comenetz, Joshua (2013). Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. *Remote Sensing of Environment, 130*, 219–232. https://doi.org/10.1016/j.rse.2012.11.022http://www.sciencedirect.com/science/article/pii/S0034425712004543.

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal, 69*(1–2), 103–117. https://doi.org/10.1007/s10708-007-

9105-9.

Bhaduri, B., Bright, E., & Rose, A. (2014). Data driven approach for high resolution population distribution and dynamics models. *Proceedings of the 2014 Winter Simulation Conference* (pp. 842–850). . Retrieved from http://dl.acm.org/citation.cfm?id=2693961.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. https://doi.org/10.1007/BF00058655.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. *International Geoscience and Remote Sensing Symposium (IGARSS),* 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393.

Bright, E. A., Rose, A. N., & Urban, M. L. (2016, January 1). LandScan 2015 high-resolution global population data set. Retrieved from https://www.osti.gov/scitech/biblio/1340997.

Center for International Earth Science Information Network (2016). CIESIN - Columbia University. *Gridded Population of the World, Version 4 (GPWv4): Population Count*Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4X63JVC (Accessed June 10, 2017).

Chave, J. (2013). The problem of pattern and scale in ecology: What have we learned in 20 years? *Ecology Letters, 16*, 4–16. https://doi.org/10.1111/ele.12048.

Dowle, M., Srinivasan, Arun, Gorecki, Jan, Chirico, Michael, Stetsenko, Pasha, Short, Tom, Lianoglou, Steve, et al. (2018). Package ‘data.table’. *Cran*.

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O., & Baptista, S. R. (2015). Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4. *Papers in Applied Geography, 1*(3), 226–234. https://doi.org/10.1080/23754931.2015.1014272.

DSD Nepal, D. S. D. of N (2015). *Health infrastructure of Nepal*.

Dumanski, J., Pettapiece, W. W., & McGregor, R. J. (1998). Relevance of scale dependent approaches for integrating biophysical and socio-economic information and development of agroecological indicators. *Soil and water quality at different scales* (pp. 13–22). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-3021-1_2.

ESRI (2016). *ArcGIS desktop: Release 10.5. Redlands CA*.

Füssel, H. M. (2007). Vulnerability: A generally applicable conceptual framework for climate change research. *Global Environmental Change, 17*(2), 155–167. https://doi.org/10.1016/j.gloenvcha.2006.05.002.

Gardner, R. H., Mime, B. T., Turner, M. G., & Neill, R. V. O. (1987). Neutral models for the analysis of broad-scale landscape pattern. *Landscape Ecology, 1*(1), 19–28. Retrieved from http://landscape.zoology.wisc.edu/People/Turner/Gardner1987NLM.pdf.

Gaughan, A. E., Stevens, F. R., Huang, Z., Nieves, J. J., Sorichetta, A., Lai, S., ... Tatem, A. J. (2016). Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Scientific Data, 3*, 160005. https://doi.org/10.1038/sdata.2016.5.

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., & Tatem, A. J. (2013). High resolution population distribution maps for southeast Asia in 2010 and 2015. *PLoS One, 8*(2), e55882. https://doi.org/10.1371/journal.pone.0055882.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*(3), 189–206. https://doi.org/10.1007/978-3-642-01976-0_10.

Goetz, J. N., Brenning, A., Petschko, H., & Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & Geosciences, 81*, 1–11. https://doi.org/10.1016/j.cageo.2015.04.007.

Griffith, D. A. (1992). Simplifying the normalizing factor in spatial autoregressions for irregular lattices. *Papers in Regional Science, 71*(1), 71–86. https://doi.org/10.1111/j.1435-5597.1992.tb01749.x.

Griffith, D. A. (2005). Effective geographic sample size in the presence of spatial autocorrelation. *Annals of the Association of American Geographers, 95*(4), 740–760. https://doi.org/10.1111/j.1467-8306.2005.00484.x.

Gustafson, E. (1998). Pattern: What is the state of the art? *Ecosystems, I,* 143–156. https://doi.org/10.1007/s100219900011.

Hahn, M. B., Riederer, A. M., & Foster, S. O. (2009). The Livelihood vulnerability index: A pragmatic approach to assessing risks from climate variability and change-A case study in Mozambique. *Global Environmental Change, 19*(1), 74–88. https://doi.org/10.1016/j.gloenvcha.2008.11.002.

Hay, S. I., Noor, A. M., Nelson, A., & Tatem, A. J. (2005). The accuracy of human population maps for public health application. *Tropical Medicine and International Health, 10*(10), 1073–1086. https://doi.org/10.1111/j.1365-3156.2005.01487.x.

Heuvelink, G. B. M. (1998). Uncertainty analysis in environmental modelling under a change of spatial scale. *Nutrient Cycling in Agroecosystems, 50*(1/3), 255–264. https://doi.org/10.1023/A:1009700614041.

Hijmans, R. J. (2016). raster: Geographic data analysis and modeling. Retrieved from https://cran.r-project.org/package=raster.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology, 25*(15), 1965–1978. https://doi.org/10.1002/joc.1276.

Hillger, D., Kopp, T., Lee, T., Lindsey, D., Seaman, C., Miller, S., ... Rink, T. (2013). First-light imagery from Suomi NPP VIIRS. *Bulletin of the American Meteorological Society, 94*(7), 1019–1029. https://doi.org/10.1175/BAMS-D-12-00097.1.

Jiang, Z., Li, Y., Shekhar, S., Rampi, L., & Knight, J. (2017). Spatial ensemble learning for heterogeneous geographic data with class ambiguity: A summary of results. *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 23). . https://doi.org/10.1145/3139958.3140044.

Jiang, Z., & Shekhar, S. (2017). *Spatial big data science.* Cham: Springer International Publishinghttps://doi.org/10.1007/978-3-319-60195-3.

Journel, A. G., & Huijbregts, C. J. (2003). *Mining geostatistics.* Blackburn Press.

King, D., Fox, D. M., Daroussin, J., Le Bissonnais, Y., & Danneels, V. (1998). Upscaling a simple erosion model from small areas to a large region. *Nutrient Cycling in Agroecosystems, 50*(1/3), 143–149. https://doi.org/10.1023/A:1009779909498.

Lehner, B., Verdin, K., & Jarvis, A. (2013). HydroSHEDS technical documentation version 1.2. *EOS Transactions, 89*(10), 26. https://doi.org/World Wildlife Fund US, Washington, DC. Available from: http://hydrosheds.cr.usgs.gov.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News, 2*(3), 18–22. Retrieved from http://cran.r-project.org/doc/Rnews/.

Liaw, A., & Wiener, M. (2015). *Package ‘ random forest ’. Breiman and Cutler's random forests for classification and regression.* (CRAN Reference Manual).

Linard, C., Gilbert, M., & Tatem, A. J. (2011). Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal.* https://doi.org/10.1007/s10708-010-9364-8.

Linard, C., & Tatem, A. J. (2012). Large-scale spatial population databases in infectious disease research. *International Journal of Health Geographics, 11*(1), 7. https://doi.org/10.1186/1476-072X-11-7.

López-Carr, D., Pricope, N. G., Aukema, J. E., Jankowska, M. M., Funk, C., Husak, G., & Michaelsen, J. (2014). A spatial analysis of population dynamics and climate change in Africa: Potential vulnerability hot spots emerge where precipitation declines and demographic pressures coincide. *Population and Environment, 35*(3), 323–339. https://doi.org/10.1007/s11111-014-0209-0.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer, 55*(1), 31–42. https://doi.org/10.1111/0033-0124.10042.

Mennis, J. (2009). Dasymetric apping for estimating population in small areas. *Geography Compass, 3*(2), 727–745. https://doi.org/10.1111/j.1749-8198.2009.00220.x.

Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science, 33*(3), 179–194. https://doi.org/10.1559/152304006779077309.

Nagle, N. N., Buttenfield, B. P., Leyk, S., & Spielman, S. (2014). Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers, 104*(1), 80–95. https://doi.org/10.1080/00045608.2013.843439.

Nieves, J. J., Stevens, F. R., Gaughan, A. E., Linard, C., Sorichetta, A., Hornby, G., ... Tatem, A. J. (2017). Examining the correlates and drivers of human population distributions across low- and middle-income countries. *Journal of the Royal Society, Interface, 14*(137), 20170401. https://doi.org/10.1098/rsif.2017.0401.

O'Neill, R. V., Gardner, R. H., & Turner, M. G. (1992). A hierarchical neutral model for landscape analysis. *Landscape Ecology, 7*(1), 55–61. https://doi.org/10.1007/BF02573957.

Pebesma, E., Bivand, R., Rowlingson, B., & Gomez-Rubio, V. (2013). *Sp: Classes and methods for spatial data.* (*URL Http://CRAN. R-Project. Org/Package= Sp, R Package Version*, 1.0-14).

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., ... Zanchetta, L. (2013). A global human settlement layer from optical HR/VHR RS data: Concept and first results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6*(5), 2102–2131. https://doi.org/10.1109/JSTARS.2013.2271445.

Pezzulo, C., Bird, T., Utazi, E. C., Sorichetta, A., Tatem, A. J., Yourkavitch, J., & Burgert-Brucker, C. R. (2016a). Geospatial modeling of child mortality across 27 countries in Sub-Saharan Africa. *DHS spatial analysis report no. 13*Rockville, Maryland, USA: ICF International. Retrieved from http://dhsprogram.com/pubs/pdf/SAR13/SAR13.pdf.

Pezzulo, C., Bird, T., Utazi, E. C., Sorichetta, A., Tatem, A. J., Yourkavitch, J., & Burgert-Brucker, C. R. (2016b). Geospatial modeling of child mortality across 27 countries in Sub-Saharan Africa. *DHS spatial analysis report no. 13*. Rockville, Maryland, USA: ICF International.

Python Software Foundation (2013). Python language reference, version 2.7. *Python software foundation*https://doi.org/https://www.python.org.

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.r-project.org.

Radcliffe, D. E., Gupte, S. M., & Box, J. E., Jr. (1998). Solute transport at the pedon and polypedon scales. *Nutrient Cycling in Agroecosystems, 50*(1/3), 77–84. https://doi.org/10.1023/A:1009703304046.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351. https://doi.org/10.2307/2087176.

Salje, H., Lessler, J., Paul, K. K., Azman, A. S., Rahman, M. W., Rahman, M., ... Cauchemez, S. (2016). How social structures, space, and behaviors shape the spread of infectious diseases using chikungunya as a case study. *Proceedings of the National Academy of Sciences of the United States of America, 113*(47), 13420–13425. https://doi.org/10.1073/pnas.1611391113.

Seligman, M. (2016). Rborist: Extensible, parallelizable implementation of the random forest algorithm. Retrieved from https://cran.r-project.org/package=Rborist.

Sorichetta, A., Bird, T. J., Ruktanonchai, N. W., zu Erbach-Schoenberg, E., Pezzulo, C., Tejedor, N., ... Tatem, A. J. (2016). Mapping internal connectivity through human migration in malaria endemic countries. *Scientific Data, 3*. https://doi.org/10.1038/sdata.2016.66 (160066).

Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). High-resolution gridded population datasets for latin America and the caribbean in 2010, 2015, and 2020. *Scientific Data, 2*, 150045. https://doi.org/10.1038/sdata.2015.45.

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), https://doi.org/10.1371/journal.pone.0107042.

Tatem, A. J. (2014). Mapping the denominator: Spatial demography in the measurement of progress. *International Health, 6*(3), 153–155. https://doi.org/10.1093/inthealth/ihu057.

Tatem, A. J., Adamo, S., Bharti, N., Burgert, C. R., Castro, M., Dorelien, A., ... Balk, D.

(2012). Mapping populations at risk: Improving spatial demographic data for infectious disease modeling and metric derivation. *Population Health Metrics, 10*(1), 8. https://doi.org/10.1186/1478-7954-10-8.

Tejedor-Garavito, N., Dlamini, N., Pindolia, D., Soble, A., Ruktanonchai, N. W., Alegana, V., ... Kunene, S. (2017). Travel patterns and demographic characteristics of malaria cases in Swaziland, 2010–2014. *Malaria Journal, 16*(1), 359. https://doi.org/10.1186/s12936-017-2004-8.

The National Research Council (2007). *Tools and methods for estimating populations at risk from natural disasters and complex humanitarian crises.* The National Academy of Sciences4.

Tiecke, T. G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., ... Dang, H.-A. H. (2017). Mapping the world population one building at a time. Retrieved from https://arxiv.org/pdf/1712.05839.pdf.

UNEP-WCMC (2010). *Data standards for the world database on protected areas, UNEP-WCMC. Database*.

Vargo, J., Habeeb, D., & Stone, B. (2013). The importance of land cover change across urban-rural typologies for climate modeling. *Journal of Environmental Management, 114*, 243–252. https://doi.org/10.1016/j.jenvman.2012.10.007.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from http://ggplot2.org.

Wu, J., Jelinski, D. E., Luck, M., & Tueller, P. T. (2000). Multiscale analysis of landscape heterogeneity: Scale variance and pattern metrics. *Annals of GIS, 6*(1), 6–19. https://doi.org/10.1080/10824000009480529.

Yuan, J. (2016). Automatic building extraction in aerial scenes using convolutional networks. *ArXiv.*. https://doi.org/10.1109/TPAMI.2017.2750680.