5-2016

Global population distributions and the environment : discerning observed global and regional patterns.

Jeremiah J. Nieves

GLOBAL POPULATION DISTRIBUTIONS AND THE ENVIRONMENT: DISCERNING OBSERVED GLOBAL AND REGIONAL PATTERNS

By

Jeremiah J. Nieves
B.S., University of Louisville, 2013

A Thesis
Submitted to the Faculty of the
College of Arts and Sciences of the University of Louisville
In Partial Fulfillment of the Requirements
for the Degree of

Master of Science
in Applied Geography

Department of Geography & Geosciences
University of Louisville
Louisville, Kentucky

May 2016

GLOBAL POPULATION DISTRIBUTIONS AND THE ENVIRONMENT:
DISCERNING OBSERVED GLOBAL AND REGIONAL PATTERNS

By

Jeremiah J. Nieves
B.S., University of Louisville, 2013

A Thesis Approved on

April 15, 2016

By the following Thesis Committee

____
Andrea E. Gaughan, PhD

____
Forrest R. Stevens, PhD

____
Catherine Linard, PhD

ACKNOWLEDGEMENTS

I would like to thank Forrest Stevens and Andrea Gaughan for their encouragement and guidance throughout this project as well as other related population research projects. I would also like to thank Amy Ninneman for her copy editing skills she graciously volunteered to the writing of this thesis. Additionally, Catherine Linard's input on the manuscript was extremely helpful in refining the results section. Further thanks is due to Graeme Hornby, Alessandro Sorichetta, and Nirav N. Patel for their assistance in wrangling the metadata of all the population models. The rest of the WorldPop team is acknowledged for their support and feedback in presentations of preliminary findings.

ABSTRACT

GLOBAL POPULATION DISTRIBUTIONS AND THE ENVIRONMENT: DISCERNING OBSERVED GLOBAL AND REGIONAL PATTERNS

Jeremiah J. Nieves

April 15, 2016

Between 1990 to 2015, numerous groups used ancillary data about the environment surrounding populations to more accurately map global populations from standard census data. No comprehensive study has been undertaken to characterize the observed relationships between population density and ancillary data. Better understanding these relationships may produce more accurate population maps, focus resources on new datasets with a high probability of modelling importance, and lead to expanded end-user applications. This study examined these relationships by extracting variable importances from 36 independently run, country-specific population models from the WorldPop project's population data. Covariate data describing urban/suburban extents were found to be the most significant predictors of population. Little difference was found in the resolution of urban/suburban data regarding their modelling importance. Further examination of the effect of different definitions of built-/urban-area, methods of quantifying input data quality, and the probability of specific variable classes as significant predictors of population is required.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Between 2015 and 2050, the U.N. (2015) estimates that the global human population will grow by 2.4 billion, with 1.3 billion and 0.9 billion being added in Africa and Asia, respectively. Most of this projected change is anticipated to occur in the least developed countries and in urbanized areas (U.N. 2014a; U.N. 2015). In this same time period, Africa, Asia, Latin America, and the Caribbean are estimated to experience the highest rates of urbanization (U.N. 2014b). As a part of this "urban transition," the majority of Africa and Asia and are experiencing large rates of internal migration, international migration, and changes in the spatial distribution of natural population growth (U.N. 2014b; U.N. 2015). And while Latin America and the Caribbean are predicted to experience decreasing urbanization rates, as was the trend through the 1990s and the early 2000s, the region is expected to have major demographic shifts. These include more than a doubling of the proportion of populations over the age of sixty years old, to 26 percent of the regional total, by 2050 (U.N. 2015). These rapidly changing magnitudes, composition, and distribution of human populations imply a continued if not increasing need for high-resolution spatially-explicit population maps which more accurately capture these changes.

Large and rapid population changes are occurring with respect to total population counts as well as with respect to internally shifting demographics, as longevity increases, high-volume international migrations continue, and high rates of urbanization proliferate (U.N. 2015, U.N. 2014b). In developing regions, such as the regions sampled for this study, these shifts in spatial population distribution and their magnitude continue to raise concerns of sustainability, infrastructure, health as related to infectious and chronic diseases, food security, and increasing energy demand at local, regional, and global scales (Cohen 2006; McGranahan et al. 2007; Stephenson, Newman and Mayhew 2010; Chongsuvivatwong et al. 2011; Madlener and Sunak 2011; Sverdlik 2011; Buhaug and Urdal 2013; Masters et al. 2013). These continued and heightened concerns regarding the implications of the rapid pace of shifting populations and demographic distributions in developing areas ensures a continued demand for high resolution gridded population maps in these regions of the world.

By understanding and clarifying the observed importance of a variety of ancillary data sources in relation to corresponding population densities, continued and future global high-resolution population mapping efforts can progress with a more complete characterization of populations and their surrounding environment. Quantifying the relative importance of ancillary data allows for population mappers to concentrate their resources on finding or developing new ancillary datasets which have the highest probability of being important when placed in a modeling framework. Clarifying the variable importance, or more specifically the non-importance of certain covariates, can lead to the formulation of a reduced covariate set for population mapping. This reduced covariate set can expand the possible end-use applications of the population data by minimizing the number of covariates that could become “circularly regressed” in subsequent regression or other statistical analyses. Moreover, by further depicting the knowledge of the relationships between the categorized ancillary datasets and population densities at global and regional scales the accuracy and precision of high resolution population mapping will be furthered.

LITERATURE REVIEW

A Review of Global Population Mapping Efforts

Since the 1990s, there have been several notable gridded population map producers which utilized a variety of statistical methods and input data to estimate population on a global or regional extent. Such efforts include the Global Rural Urban Mapping Project (GRUMP), Gridded Population of the World (GPW), LandScan, the United Nations Environment Programme (UNEP), and WorldPop (formerly known as AfriPop and AsiaPop) (Dobson et al. 2000; UNEP 2004; Balk and Yetman 2004; Balk et al. 2006; Bhaduri et al. 2007; Cheriyadat et al. 2007; Linard et al. 2010; CIESIN 2011). GRUMP version 1, GPW version 3, and the beta of GPW version 4 are freely available as are the UNEP gridded datasets for Africa, Asia, and Latin America (Balk and Yetman 2004; UNEP 2004; CIESIN 2011; Doxsey-Whitfield et al. 2015). WorldPop currently hosts the datasets modeled by AfriPop and AsiaPop, many of which have been updated with new data or methods since those projects merged in 2013 into what is now WorldPop (Stevens et al. 2015; WorldPop 2015c). Alternatively, the LandScan project produces commercially available data and has the advantage of being updated on an annual basis (Dobson et al. 2000; Bhaduri et al. 2007). Another notable difference is that LandScan maps the “ambient” population over a 24-hour period as opposed to the traditional night-time residential population locations captured by a census (Dobson et al. 2000; Bhaduri et al. 2007). These population maps span a variety of spatial and temporal resolutions all of which are detailed in Table 1 and are chronologically presented in Figure 1.

Table 1. Contemporary population datasets and supplemental derived products

Population Map	Spatial Resolution (approx. at equator)	Spatial Extent	Temporal Resolution	Temporal Extent	Updating Period	Supplemental Data Produced
LandScan	30 arc sec (1km)	Global	Ambient Population (24 hour average)	1998 - 2012	Annual	---
GPW v3	2.5 arc min (5km)	Global	Single Time Point	1990, 1995, 2000 & 2015, 2020 projections	Intermittent	---
GRUMP v1	30 arc sec (1km)	Global	Single Time Point	1990, 1995, 2000	Intermittent	Urban-Rural Classification
UNEP	1° (111km)	Global	Single Time Point	1990	None	---
	2.5 arc min (5km)	Africa	Single Time Point	1960, ’70, ’80, ’90, 2000	None	---
	2.5 arc min (5km)	Asia	Single Time Point	1995	None	---
	2.5 arc min (5km)	Latin America and Caribbean	Single Time Point	1960, ’70, ’80, ’90, 2000	None	---
WorldPop	3 arc sec (100m)	South and Central America, Parts of Africa and Asia	Single Time Point	Input census year & 2010, 2015, 2020 back/forward projections	As data becomes available	Small area demographic estimations and select small area characteristics related to specific health outcomes

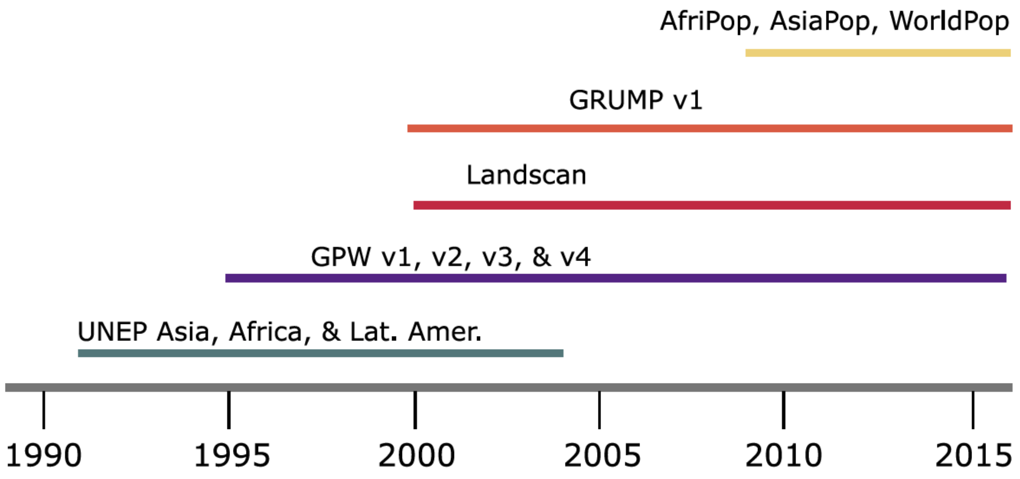


Figure 1. Timeline of global population mapping projects

Global Population Mapping Methods

GPW utilizes an areal weighting technique, in which population counts, obtained from census data, for a given areal census unit are broken into smaller, spatially coincident areas of population (Tobler et al. 1997). This is carried out by assuming uniform distribution of a population across the source area and deriving the population counts of the smaller target areas based upon the proportion of their area to the source unit area; the summed population of the target areas equals the population of the source area (Flowerdew, Green, & Kehris 1991). This technique was improved upon and greater spatial variance of populations within census units was introduced in GRUMP through the integration of urban-rural designations derived from lights-at-night satellite data, city point, and other geographic datasets (Balk and Yetman 2004; Balk et al. 2006; CIESIN 2011). Further increases accuracy evolved within UNEP, LandScan, and WorldPop projects fusing dasymetric mapping with statistical techniques to estimate population locations and densities (Dobson et al. 2000; UNEP 2004; Bhaduri et al. 2007; Stevens et al. 2015).

Dasymetric mapping redistributes values from coarse spatial resolution to a finer spatial resolution within a given spatial unit using either uniform, areal, or non-uniform spatial weights. These non-uniform weights may be determined by statistical relationships with independent, ancillary datasets which provide correlational information about population density (Mennis and Hultgren 2006). The relationships between spatially coincident populations and ancillary data are typically determined *a priori*, through expert knowledge, or by the distributions of the covariates in relation to the output of interest as assessed through statistical methods or machine learning algorithms (Wright 1936; Eicher and Brewer 2001; Langford et al. 1991; Mennis and Hultgren 2006; Stevens 2015). Between 1991 and present day, the advancement of statistical techniques as applied to dasymetric methods, access to ancillary datasets, GIS and remote sensing, and the availability of processing power have paralleled an increase in accuracy of gridded population products.

Global Population Mapping Covariates

Some ancillary data sets chosen for disaggregating populations from census units represent phenomena known to be related to population. For instance, it has been known that humans tend to modify their environment, specifically land cover, in manners that differentiate it from the surrounding landscape (Meyer and Turner 1992). This is especially true for urbanized areas and areas of mono-agriculture although the exact direction and magnitude of the relationship may vary with locale (Ramankutty, Foley and Olejniczak 2002; Pozzi and Small 2005) . Other examples include the increased probability of settlements occurring within a specific distance of rivers and coasts as well as the phenomena of populated settlements giving off light, from campfires, street lights, etc., that are visible from satellites at night (Elvidge et al. 2001; Small and Nicholls 2003; McGranahan, Balk and Anderson 2007). Other ancillary data included have less clear relationships such as transportation networks, elevation, impervious surface cover, points of interest, and environmental factors. Some of these relationships vary widely within a given country (e.g. impervious surface may be indicative of population in one area and non-populated industry in another area) and or between countries (e.g. distance to railways may be important in Myanmar and not in Thailand). Making the choice of covariates in model selection an important process.

Research Question and Hypotheses

Despite the variety of analytical approaches and ancillary data used in the creation of high-resolution gridded population maps, and the extensive application of these maps over the previous decade, there is a lack of comprehensive studies on the observed spatial relationships between population densities, the covariates they are associated with, and the ancillary datasets that represent the covariates. That is, there has been no meta-analysis of the relative importance and effectiveness, of ancillary datasets in estimating the spatial distribution of populations at either country or regional levels. At most, basic within-country analyses have been undertaken in the course of validation or accuracy assessment (Gaughan et al. 2013; Stevens et al. 2015).

There are four primary questions that I attempt to address through this research: (1) What types of geospatial ancillary data are the most important for dasymetrically mapping populations at a global and regional scale? (2) What are the differences in the patterns of importance of covariate categories between and within regions? and (3) How important are built-area/urban extent data to population distribution modeling and is spatial resolution significant in determining this? Corresponding hypotheses and scale of analyses to investigate these questions are shown with expected results in Table 2. For the purposes of this study “Southeast Asia” refers to the proper Southeast Asia region with China and Nepal included.

Table 2. Hypotheses, corresponding research question addressed, scale and unit of analysis, and expected results

Hypothesis	Research Question	Scale [Unit] of Analysis	Expected Results
Certain classes of ancillary data will be significantly more important in explaining observed population density as measured by Percent Increase in MSE	1	Global [Covariate Classes]	Urban/suburban covariate categories will be the most important followed by transportation and facilities/services covariates
	2	Inter-Regional [Regions]	Urban/suburban covariate categories will be the most important followed by transportation and facilities/services covariates. The magnitude of these importances will be sig. diff. between regions.
	2	Intra-Regional [Covariate Classes]	Urban/suburban covariate categories will be the most important followed by transportation and facilities/services covariates. There will be some countries within a given region that will be significantly different.
Urban/suburban extent and urban/suburban extent proxies will significantly vary in importance inter- and intra- regionally largely due to their data source/resolution	3	Globally [Countries]	Urban variable resolution will be significantly different in importance with the higher resolution or vector data sets being more important than the lower resolution variables.
	3	Intra-Regional [Regions]	Urban/suburban covariate categories will be more important in Central America and the Caribbean, South America, and Southeast Asia as compared to Africa primarily because of the 30m urban variables that are derived from their land cover datasets, whereas Africa has 300m.
Certain classes of ancillary data will be significantly more important in explaining observed population density as measured by the within-country weighted rank of variable importance	1	Global [Countries]	Urban/suburban covariate categories will be the most important followed by transportation and facilities/services covariates
	2	Inter-Regional [Regions]	Urban/suburban covariate categories will be the most important followed by transportation and facilities/services covariates. The magnitude of these importances will be sig. diff. between regions with it being most important in Southeast Asia.
	2	Intra-Regional [Countries]	Urban/suburban covariate categories will be the most important followed by transportation and facilities/services covariates. There will be some countries within a given region that will be sig. diff. from the category's median rank in the region.

DATA AND METHODS

The WorldPop Project's methods and data sources are open source and transparent resulting in the utilization of the population maps by non-profit groups, governmental entities, non-governmental organizations, and academic researchers in a variety of applications (World Pop 2015a). Such applications include health prospect and risk assessment, guiding of vaccination and health intervention campaigns, natural disaster impact assessment and relief coordination in the 2015 Nepal earthquake and 2015 Myanmar floods, and in response to the Ebola epidemic in West Africa (WorldPop 2014; Tatem et al. 2014; W.H.O. 2014; UNFPA 2014; WorldPop 2015a; WorldPop 2015b; WorldPop 2015c; Alegana et al. 2015; Bharti et al. 2015). The methods used are a two-step process involving the use of random forest regressions to determine appropriate models and covariates to estimate population density and then the application of intelligent dasymetric mapping, guided by predicted per-pixel population density produced by the random forest, to redistribute census population counts across space at the pixel level.

Random Forests – An Ensemble Learning Method

Random forests (RFs) are a non-parametric and non-linear statistical method which falls within a category of machine learning methods known as "ensemble methods." Multiple decision trees, that on their own are considered "weak learners," i.e. they are only slightly correlated with the training data, are combined in order to create a "strong learner." Ensemble methods utilize a training data set to build a number of models (e.g. decision trees) by using an *allocation function* to determine how much of the training data each model receives. These multiple models are then combined through a *combination function* which determines how best to resolve disagreements amongst the models' predictions (e.g. through voting, weighting, etc.). The output is a single *ensemble model*. Ensemble methods differ primarily in their allocation functions and their combination functions as well as the method used to create the multiple models (Dietterich 2000). The benefit of ensemble methods is that generalizability is increased, performance on extremely large or small datasets is improved, and the ability of the method to "understand" or model difficult learning tasks is more nuanced and effective. Additionally, the ensemble model produced is able to synthesize or predict data from very specific and distinct domains. A generalized layout of ensemble methods is shown in Figure 2.

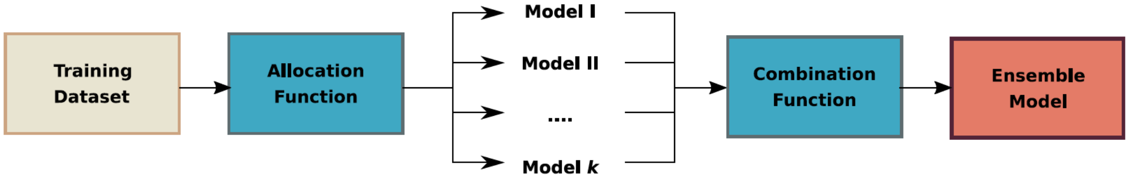


Figure 2. General schematic of ensemble methods

RFs independently generates *k* number unpruned of decision trees using "bagging," in which two-thirds of the training data set is boot-strap sampled with replacement (Breiman 1996; Breiman 2001). At each decision node in a given decision tree, the data is split into two subsets based upon a random selection of *m* attributes as the split decision criteria, with the two resulting subsets being as homogenous in their attributes as possible (Breiman 2001). Once a decision tree has been grown, the remaining one-third of the training data which the tree was not grown upon, known as the "Out-of-Bag" (OOB) data, has the decision tree applied to it and the accuracy of the decision tree in classifying or regressing that data, as measured by the mean squared error (MSE), is stored as the OOB error for that tree (Breiman 2001). The prediction error of the entire RF model can be estimated by averaging the OOB error of all the constituent trees (Breiman 2001). Additionally, the OOB error can be used for estimating covariate importance by replacing a given covariates OOB data with random noise and calculating the percent increase in the OOB error of the RF model (Breiman 2001). The overall variance explained by the model is equivalent to one minus the mean squared residuals as shown in Equation 1 where $\hat{y}_i^{OOB}$ is the average of the OOB predictions for the *i*th observation and $\hat{\sigma}_y^2$ is calculated with *n* as the divisor as opposed to *n* – 1 (Liaw and Wiener 2002).

$$Variance\ Explained = \ 1 - n^{-1} * \sum_{1}^{n} (y_i - \hat{y}_i^{OOB})^2 \Big/ \hat{\sigma}_y^2 \qquad (1)$$

The output of all trees can be consolidated by majority vote (i.e. the mode of the outputs) or by calculating the average of the outputs, for RFs used for classification or regression, respectively (Breiman 2001). A general schematic of the RF process is given in Figure 3.

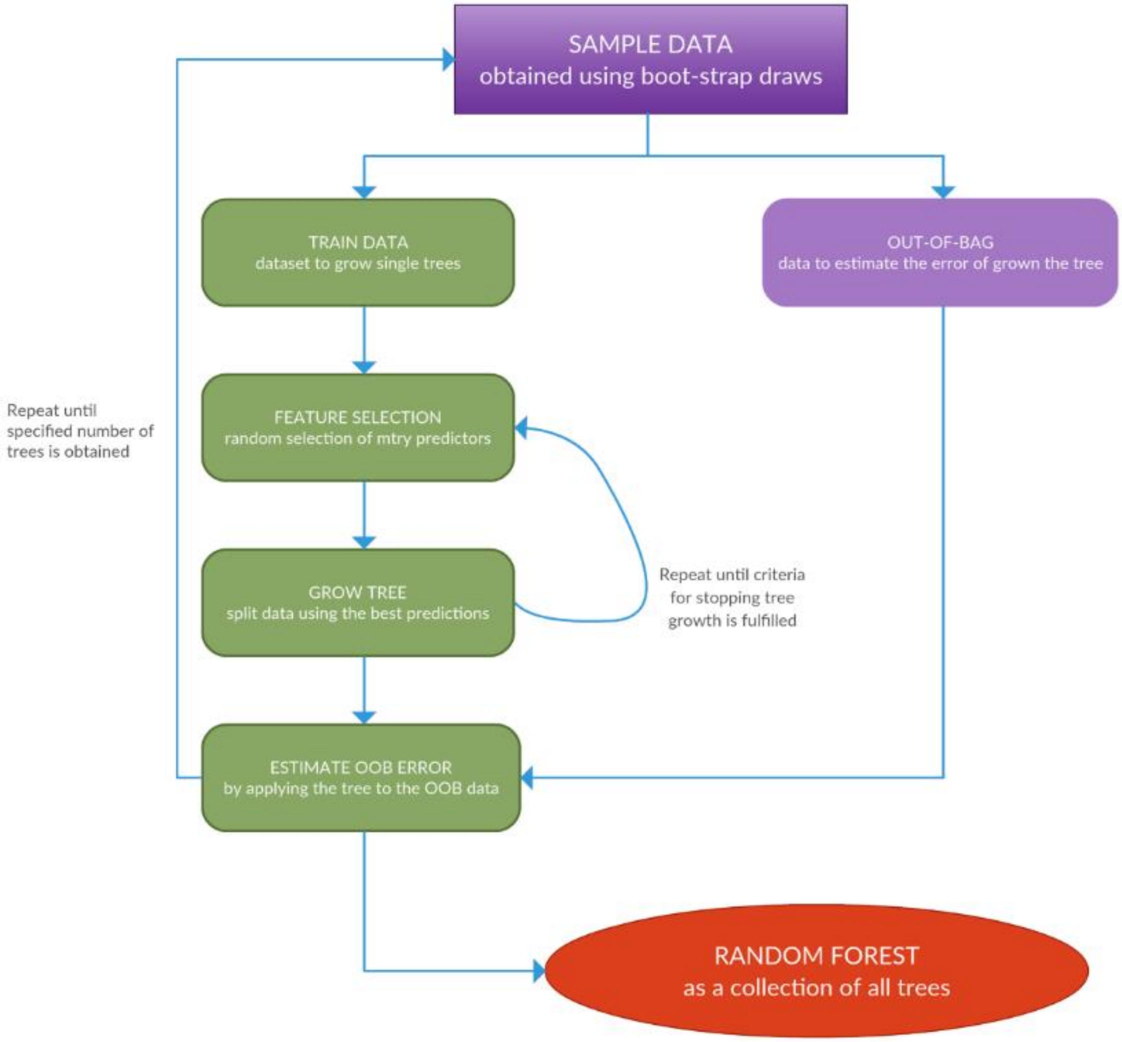

SAMPLE DATA
obtained using boot-strap draws
TRAIN DATA
dataset to grow single trees
OUT-OF-BAG
data to estimate the error of grown the tree
Repeat until
specified number of
trees is obtained
FEATURE SELECTION
random selection of mtry predictors
GROW TREE
split data using the best predictions
Repeat until criteria
for stopping tree
growth is fulfilled
ESTIMATE OOB ERROR
by applying the tree to the OOB data
RANDOM FOREST
as a collection of all trees

Figure 3. General process of building a random forest.

Compared to other ensemble methods RFs are robust to noise, small sample sizes, and over-fitting, yet they need little in the way of parameter specifications (Feller 1968; Breiman 2001; Liaw and Wierner 2002; Briem et al. 2002; Pal and Mather 2003; Chan and Paelickx 2008; Rodriguez-Galiano et al. 2012). There are three primary parameters in constructing a RF: (1) *m* number of covariates to be randomly selected at each node, (2) *k* number of trees in the forest, and, (3) the number of observations allowed in the terminal nodes of each decision tree (Liaw and Wiener 2002). The optimal number of covariates to be randomly selected and the number of observations allowed can be automatically selected, based upon minimizing the OOB Error, by using the *tuneRF* function in the R package *randomForest* (Liaw and Wiener 2002).

WorldPop Random Forest-Based Dasymetric Population Mapping

WorldPop uses a RF regression model and dasymetric mapping methods in a three step process to estimate a population layer from input census and covariate data. The steps are as follow: (1) Covariate selection for the RF model, (3) the fitting of the RF model and creation of a population density weighting layer from the created RF model, and, (3) the dasymetric redistribution of population counts from census-based administrative units to grid cells using the population density weighting layer (Stevens et al. 2015). Data input to a RF model varies on a country-by-country basis with high-resolution country specific datasets being used over coarser resolution default datasets when available. Typical ancillary datasets include land cover, elevation, transportation network, climatological, hydrological, and settlement data (Gaughan et al. 2013; Stevens et al. 2015; Sorichetta et al. 2015). All input data is projected, rasterized (if applicable) and resampled to 100m, using techniques appropriate for the given dataset (Stevens et al. 2015).

The covariate selection occurs by fitting a model at the administrative unit level of the input census data using all available covariates with the log-transformed population density of each administrative unit as the outcome of interest (Stevens et al. 2015). An iterative covariate selection process then occurs based upon the observed covariate importance, as derived from the OOB error, with covariates exhibiting zero percent increase in the model's MSE being removed prior to refitting the model (Stevens et al. 2015). The final covariate selection is determined when only covariates with positive percent increase in MSE (Per.Inc.MSE) remain (Stevens et al 2015).

The RF models are then fit by growing 500 trees, using the previously determined covariates, and setting the number of terminal node observations to one or the number of administrative units divided by 1000 and rounded to the nearest whole number, should there be greater than 1000 administrative units in the input data (Stevens et al 2015). Once grown, this RF model was used to predict the log population density, later back transformed, of every given grid cell in the model area with the average prediction of all trees being assigned to a given grid cell (Stevens et al. 2015).

This population density weighting layer is back-transformed and used to dasymetrically redistribute the input census population count values for the year of the census data (Stevens et al. 2015). This redistributed population is then projected to 2010, 2015, and 2020 using country-specific urban and rural population growth rates given by the 2014 UN projected growth estimates (U.N. 2014a; Stevens et al. 2015). The final 100m x 100m gridded population maps are output in un-projected format as people per pixel and in projected format as people per hectare (Stevens et al. 2015). In a dasymetric mapping context, when the magnitude, direction, and structure of the covariate relationships are quantified through the use of a statistical method on the distribution of the covariates in relation to the outcome (i.e. population density), the method chosen (e.g. multiple linear regression as opposed to a regression tree) can significantly determine the resulting variable importances and relationships observed between the covariates and between the covariates and the outcome.

Sampled Countries and Data

For this investigation, I sampled countries from four primary regions of the world where WorldPop has created population datasets: Africa (AFR), Central America and the Caribbean (CAC), South America (SAM), and Southeast Asia (SEA). The sampled countries within these regions, shown in Figure 4, were modeled based upon census data from varying years and the best available covariate data at the time of modelling, shown in Table 3. These regions were selected because of their continued and rapidly growing importance in relation to world population (U.N. 2014a, 2014b).

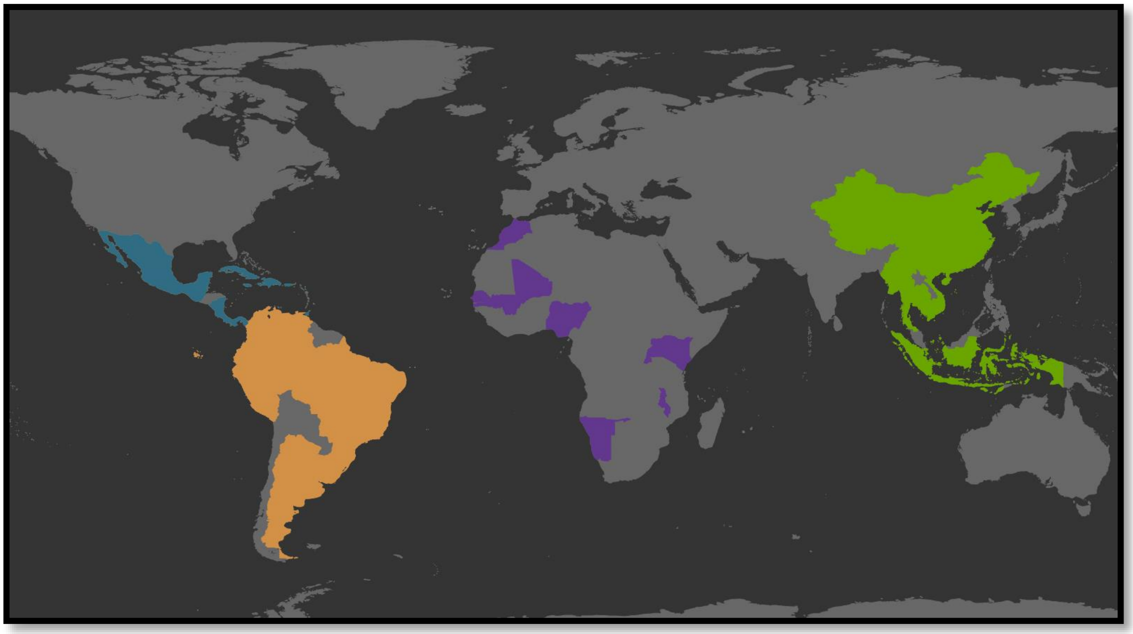


Figure 4. Sampled countries in AFR (purple), CAC (blue), SAM (orange), and SEA (green).

It is not simply the magnitude of these population changes, but the rate at which they are changing. Medium-variant projection shows Africa having an annual rate of population change peaking around 2.5 percent and decreasing to approximately 1.8 percent in 2050 (U.N. 2015). Similarly Asia and Latin America and the Caribbean peak at approximately 1.2 and 1.3 percent in 2015 and both decrease to about 0.3 percent in 2050 (U.N. 2015).

Census data is attributed to irregularly shaped polygons known as administrative units which have hierarchical classifications by “level,” which are a function of their nested subdivision and relative sizes to other levels within each country. Because these levels are not a function of their absolute size and spatial configuration, administrative units of the same level are not comparable across countries on the basis of the spatial resolution of the census data. To mitigate this, I adopted the average spatial resolution (ASR, in km^2) measure by Tobler et al. (1997) which takes the square root of the average area encompassed by a given country's administrative units, shown in Equation 2. The ASR can be thought of as a polygonal equivalent of the resolution of a pixel in a standard uniform grid, but is not directly comparable. The ASR of each country's input census data is shown in Table 3.

$$ASR = \sqrt{\frac{\sum(admin\ unit's\ area)}{N\ admin\ units\ in\ country}} \qquad (2)$$

Table 3. Sampled countries and selected characteristics including the variance explained by the country specific random forest model

Country	ISO	Region	Census Year (adm. lvl.)	N Admin Units	ASR* (km^2)	Variance Explained
Kenya	KEN	AFR	1999 (5)	6606	9	83%
Morocco	MAR	AFR	2004 (4)	1497	16	80%
Mali	MLI	AFR	2009 (4)	687	43	85%
Malawi	MWI	AFR	2008 (2)	12557	22	79%
Namibia	NAM	AFR	2011 (2)	5475	12.28	96%
Nigeria	NGA	AFR	2006 (2)	774	34	88%
Rwanda	RWA	AFR	2002 (4)	9183	1.68	69%
Senegal	SEN	AFR	2009 (4)	331	24	91.68%
Uganda	UGA	AFR	2002 (4)	5018	7	85%
Antigua and Barbuda	ATG	CAC	2011 (1)	7	7.4	86%
Belize	BLZ	CAC	2010 (1)	16	37.0	79%
Bolivia	BOL	CAC	2012 (2)	112	97.7	65%
Costa Rica	CRI	CAC	2011 (3)	469	10.4	92%
Cuba	CUB	CAC	2012 (2)	168	25.6	82%
Dominican Republic	DOM	CAC	2010 (3)	155	17.6	86%
Guatemala	GTM	CAC	2012 (2)	333	18.0	80%
Haiti	HTI	CAC	2009 (4)	570	6.9	84%
Jamaica	JAM	CAC	2011 (1)	14	28.0	86%
Mexico	MEX	CAC	2010 (2)	2456	28.0	92%
Nicaragua	NIC	CAC	2012 (3)	137	29.4	79%
Panama	PAN	CAC	2010 (2)	74	31.04	74%
Puerto Rico	PRI	CAC	2010 (1)	78	13.3	74%
Trinidad and Tobago	TTI	CAC	2011 (1)	14	19.1	86%
Argentina	ARG	SAM	2010 (2)	526	73.0	88%
Brazil	BRA	SAM	2010 (4)	5565	5.1	84%
Columbia	COL	SAM	2013 (4)	1115	32.0	84%
Ecuador	ECU	SAM	2010 (4)	978	16.2	82%
Peru	PER	SAM	2012 (2)	194	81.7	63%

Country	ISO	Region	Census Year (adm. lvl.)	N Admin Units	ASR* (km^2)	Variance Explained
Uruguay	URY	SAM	2011 (1)	19	96.0	91%
Venezuela	VEN	SAM	2011 (2)	339	51.6	71%
Cambodia	KHM	SEA	2008 (3)	1621	10.51	92%
China	CHN	SEA	2010 (4)	2922	57.28	95%
Indonesia	IND	SEA	2010 (4)	79277	4.91	81%
Myanmar	MMR	SEA	2014 (3)	326	45.29	94%
Nepal	NEP	SEA	2011 (4)	3973	6.08	92%
Thailand	THA	SEA	2010 (3)	7416	23.67	88%
Vietnam	VNM	SEA	2010 (3)	688	21.85	93%

* ASR values for CAC and SAM countries obtained from Sorichetta et al. (2015)

Rather than attempt to standardize the input covariates between countries, WorldPop has utilized the most contemporary and available datasets on a country-by-country basis to produce the population maps at nominal 100m spatial resolutions. See Stevens et al. (2015) for a typical set of ancillary data included in a given model. In some cases, where there is a lack of strong input data, a country's model can be parameterized partially on a neighboring country, however this further obfuscates the already unintuitive relationships between population density and the supporting covariates as determined by the RF model (Stevens et al. 2015). Accordingly, no countries that were parametrized on neighboring countries were included in the sample for this study.

For every WorldPop model run, metadata files containing information about the Random Forest model settings, input covariates and their importance, metadata on the input covariate datasets themselves, and the general results of the Random Forest model are output to Random Forest summaries. Due to this variability of input datasets, I extracted a variety of information, detailed in Table 4, from those RData files and examined the input covariates for all sampled countries to create the general covariates classification groups shown in Figure 5. The primary purpose of this classification system was to create some level of standardization of the covariate to be able to perform comparisons between the country models. Prior to analysis, all covariates were reclassified using the classification scheme in Figure 5.

Table 4. Information extracted from metadata files

Information	Covariate Name[*]	Level of Measurement	Description/Example
Variable Name	VAR.NAME	Nominal	*urb_dst*
Variable Classification Group	VAR.CLASS	Nominal	Aggregated variable class (See Figure 2)
Variable Percent Inc. MSE	PER.INC.MSE	Ratio	Percent Increase in MSE when covariate is removed
Variable Inc. Node Purity	INC.NODE.PURITY	Ratio	Percent purity of the variable nodes
Variable Type	VAR.TYPE	Nominal	Raw values or derived (distance, proportion, etc.)
Variable Format	FORMAT	Nominal	Raster, polygon, point, etc.
If Variable is Used By Default	DEFAULT	Logical Binary	True/False
Country Name	ISO	Nominal	Rwanda
Number of Nodes	NRNODES	Ratio	Number of nodes used in random forest regression
Variable Measure Type	MSEAURE.TYPE	Ordinal	Ratio, Nominal, etc.
Year of Census Data	CENSUS.YEAR	Ratio	1999, 2000, etc.
Region of Modelled Country	REGION	Nominal	AFR, CAC, SAM, SEA
Total Variance Explained	VAR.EXP	Ratio	Total variance explained by model

[*] These are the naming conventions utilized in the coding scripts

To incorporate a measure of variable class importance while accounting for the frequency with which those classes are not included in the final model selection, I created a “zero-inflated” variable importance dataset. This dataset included the importance of the final covariate selection, Per.Inc.MSE calculated by the random forest, and included those excluded variables by giving them a Per.Inc.MSE which I assigned a value of zero. Additionally, the Landcover No Data covariate class was used as a control to test all other covariate classes’ importance for significant difference from no data.

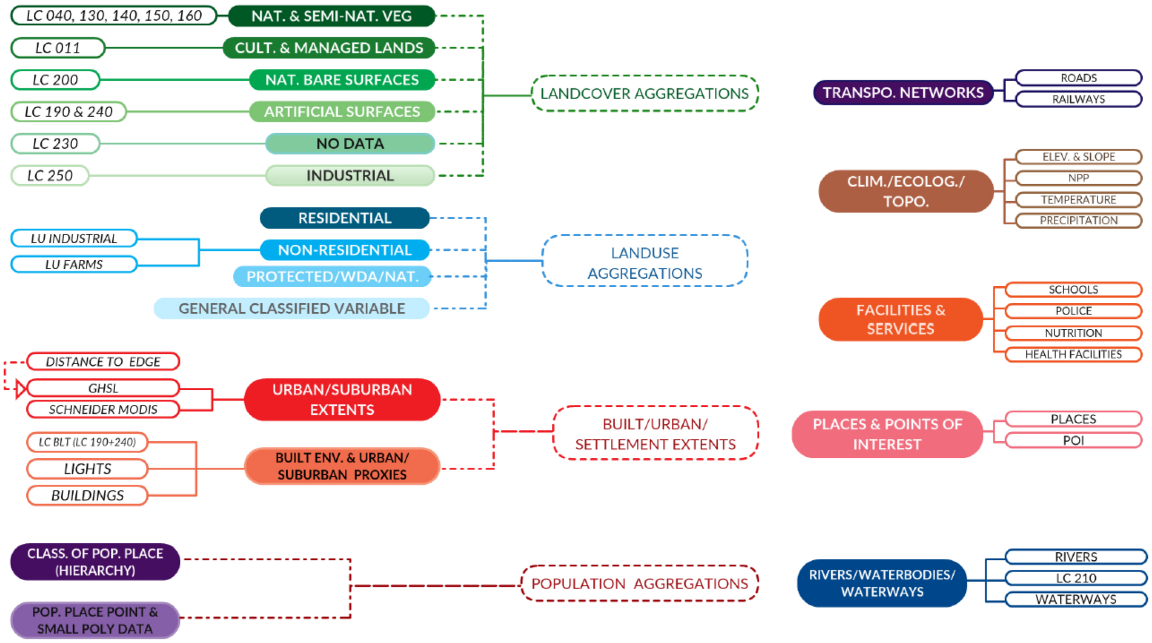


Figure 5. Covariate reclassification scheme utilized in analyses; constituent covariates are solid polygons with no fill, classes utilized in the analyses are solid, filled polygons, and the dotted polygons are the larger conceptual aggregations that guided aggregation decisions.

Analysis

This research followed the general framework of a meta-analysis. However, in the literature there are no comparable meta-analyses where individual model runs take the place of individual manuscripts within the meta-analytical framework. From these independent model runs of countries, I synthesized more generalized knowledge on the relative importance of various covariates in dasymetrically predicting population densities at high resolution.

All analysis, data extraction, and reclassification was performed in the R Statistical Environment, version 3.2.2, with α = 0.05 significance levels and appropriate corrections for multiple outcomes where indicated (R Core Team 2015). Data extraction and management utilized the “dplyr” and the “tidyr” packages in R (Wickham and

Francois 2015, Wickham 2015). All data visualizations were created using the “ggplot2” and the “RColorBrewer” packages R or the default functions in R (Wickham 2009; Neuwirth 2014). Kruskal-Wallis tests and post-hoc Dunn tests were performed using the PMCMR package (Pohlert 2016). All mapping was performed in ArcGIS 10.2 (ESRI 2013).

For all hypotheses, I calculated standard summary statistics of variable importance measures for each variable classification group. To assist in interpretation of results, I also calculated descriptive statistics for the number of administrative units by region, extracted the total variance explained by the model, and calculated the average spatial resolution across all countries (i.e. “globally”) as shown in Table 2. Lastly, I created tables of the proportion of final models that a covariate class was included in. To facilitate presentation of methods, results, and discussion of results, I used the following subheadings corresponding to the first, second, and third hypotheses presented in Table 2: Covariate Class Importance: Per.Inc.MSE; Importance of Urban Covariate Classes; and Covariate Class Importance: Weighted Rank.

Variable Class Importance: Per.Inc.MSE

To examine potential significant differences in covariate class importance as measured by Per.Inc.MSE, I utilized both analytical and graphical methods. I created boxplots of the distributions of variable class importances, both for the zero-inflated and non-zero-inflated importance datasets. I also created a line and dot plot showing the median Per.Inc.MSE values and interquartile range (IQR) of Per.IncMSE values for each covariate class grouped by region in order to examine overall regional covariate class importance patterns.

Given the non-normal nature of the variable importance data, I used the non-parametric form of Kruskal-Wallis tests to test for significant differences between covariate classes across all countries (Kruskal & Wallis 1952). The inter-regional analyses were of a hierarchical nature using data subsets of a given covariate category and using the region category as the grouping variables, but still using the Kruskal-Wallis test (Kruskal & Wallis 1952; Rosner 2011). The intra-regional analyses subset the data to a given region and a given variable class then used a Kruskal-Wallis test to determine if significant differences in importances for the given covariate class existed between countries of the same region (Kruskal & Wallis 1952). If any of the Kruskal-Wallis tests were significant they were followed up with post-hoc Dunn tests using Holm's correction for multiple outcomes (Dunn 1964; Holm 1979).

Importance of Urban Variable Classes

To examine the potential role of data resolution on the observed importance of urban related covariate datasets in predicting population density, I subset the data to include observations that were classified as the variable classes "Urban/Suburban Extents" and "Built Env. & Urban/Suburban Proxies." I extracted the corresponding spatial resolutions of each observation's source dataset. Given the non-normal distribution of this importance data, as measured by the weighted importance rank, I used a Kruskal-Wallis test to determine if there were significant differences between the different spatial resolutions across all countries (Kruskal & Wallis 1952). I also used a Kruskal-Wallis test to determine if there were significant differences between the different spatial resolutions for each given region (Kruskal & Wallis 1952). If any of the Kruskal-Wallis tests were significant they were followed up with post-hoc Dunn tests using Holm's correction for multiple outcomes (Dunn 1964; Holm 1979).

Variable Class Importance: Weighted Rank

To account for the differing number of total covariates in each country's model I calculated a weighted importance rank. Within each country, covariates were ranked according to descending Per.Inc.MSE and then weighted by the total number of covariates in the final model for a given country, as displayed in Equation 3.

$$Weighted\ Importance\ Rank = \frac{Within\ Country\ Ranked\ Importance}{Total\ Number\ of\ Covariates\ in\ Country\ Model} \tag{3}$$

Statistical testing was identical to the procedures used to examine covariate class importance as measured by MSE, with the primary procedural difference being that the hypotheses were interrogated by examining the weighted rank importance of covariate classes. Additionally, using the weighted rank importance, graphical outputs were constructed similar to those created when examining covariate class importance as measured by Per.Inc.MSE.

RESULTS

I observed consistent patterns of strong importance of the Urban/Suburban Extents, Built Env. & Urban/Surburban Proxies, and Clim./Ecolog./Topo. variable classes at both the global and regional levels. Globally and within any regions, there was no significant difference in observed importance between the Urban/Suburban Extents and the Built Env. & Urban/Suburban Proxies variable classes. As expected (Table 2), there were notable variations and significant differences in variable class importance across regions, but not necessarily between all regions. For several variable classes, the distributions of variable importance showed similar distributions between AFR and SEA and similar distributions between CAC and SAM, but significant differences between those two similar regional groups (i.e. AFR or SEA vs. CAC or SAM). No consistent, significant intra-regional differences were found across any of the variable classes.

Regarding the more general descriptive statistics, similar grouping of values for the number of administrative units can be seen, in Table 5, for AFR and SEA as well as for CAC and SAM. Also notable is the large difference in the average and median ASR for SAM as compared to other regions.

Table 5. Descriptive statistics of admin units input to model and the ASR of those units, by region

Region	Average Number of Admin. Units	Median Number of Admin. Units	Std. Dev. of Admin Units	Average ASR	Median ASR	Std. Dev. of ASR
AFR	4680.88	5018	4286.97	18.77	16	13.38
SEA	13746.14	2922	28996.19	24.22	21.85	29.20
CAC	346.50	107.50	636.70	26.38	22.35	13.53
SAM	1020.11	339	1944.90	50.80	51.60	14.78

The rate of inclusion of variable classes in the final population models, across all countries, are in Table 6. The five covariate classes with the highest rates of inclusion, were: Clim./Ecolog./Topo. (87%), Built Env. & Urban/Suburban Proxies (53%), LC Cult & Managed (57%), Urban/Suburban Extents (53%), and LU Protected (51%). Variable classes with the lowest rates of inclusion, excluding LC No Data, include: Rivers/Waterbodies/Waterways (37%), Facilities & Services (38%), and LC Nat. & Semi. Nat. Veg. (40%).

Table 6. Rate of inclusion for each variable class across all countries' final models.

Covariate Class	Rate of Inclusion
Clim./Ecolog./Topo.	87.22%
Built Env. & Urban/Suburban Proxies	58.64%
LC Cult. & Managed	57.89%
Urban/Suburban Extents	53.85%
LU Protected	51.38%
Transportation Network	44.52%
Places & POI	44.09%
LU Non-Residential	43.90%
LU Gen. Class. Var.	43.48%
Pop. Place & Small Poly. Data	42.86%
LC Nat. Bare Surfaces	42.11%
Class of Pop. Place	41.61%
LU Residential	41.18%
LC Nat. & Semi. Nat. Veg.	40.63%
Facilities & Services	38.79%
Rivers/Waterbodies/Waterways	37.73%
LC No Data	0.00%

Variable Class Importance: Per.Inc.MSE

Zero Inflated Importances

Globally, when including variables excluded from the final model of a given country as an “observed” value of zero Per.Inc.MSE, there are only a few variable classes that have median Per.Inc.MSE values that are above zero. The covariate classes that have median Per.Inc.MSE values above zero are most frequently considered important to explaining variation in population density within a country’s model. This approximately follows the line-up of the variable classes with the highest rates of inclusion shown in Table 6. The global “zero-inflated” importances are shown as variable-class-specific boxplots in Figure 6. Figure 6 presents the *distribution* of observed variable class importances “penalized” by the frequency a variable of a given class when it was not found to be a completely uncorrelated predictor of population density, and was therefore not included in the given country’s final population model. The five variable classes for predicting population density that have non-zero medians, from highest to lowest, are Clim./Ecolog./Topo. (9.93%), Built Env. & Urban/Suburban Proxies (3.37%), LC Cult. & Managed Lands (2.37%), Urban/Suburban Extents (1.96%), and LU Protected/WDA/Nat. (0.30%).

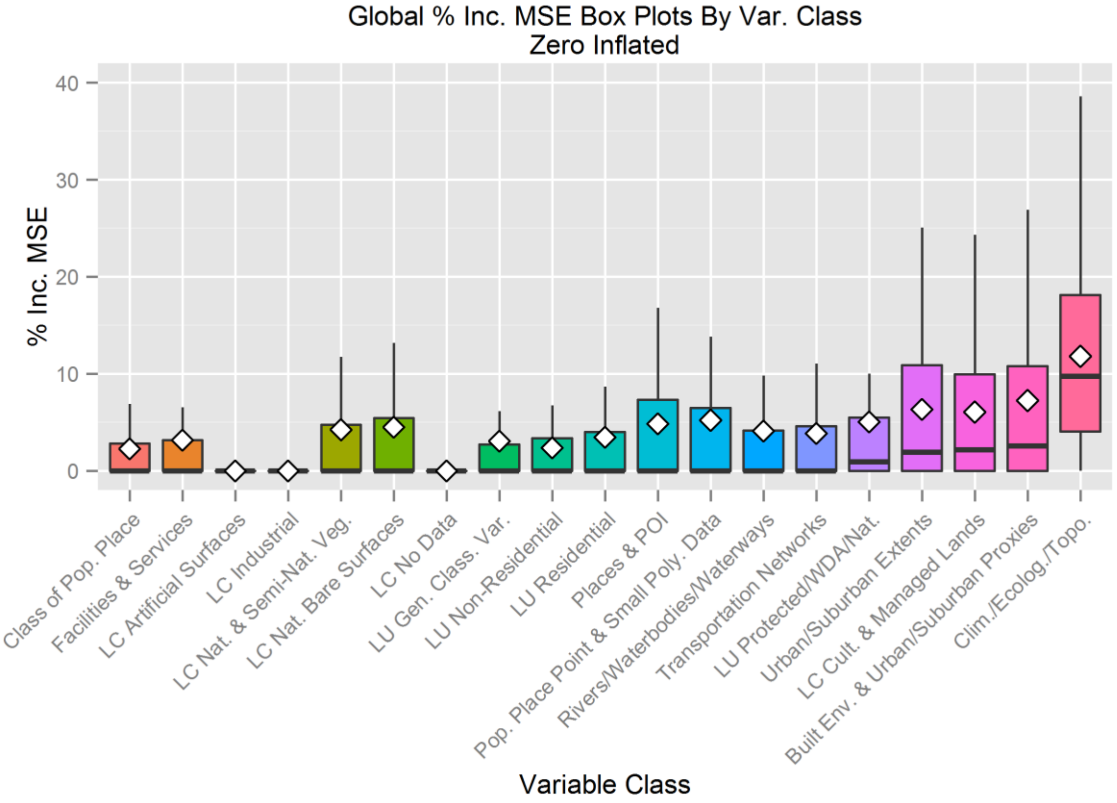


Figure 6. “Zero Inflated” values of the percent increase in the mean squared error (MSE) of each covariate class. The mean is represented by a white diamond, the median is represented by the black bar, and the whiskers represent the max and min value within 1.5 * Interquartile Range (IQR).

I investigated if a given variable class was significantly different from the LC No Data class. That is to say, I tested if the covariate classes were significantly different from no data at all. The results from comparing all covariate classes, globally, against LC No Data are shown in Table 7 and I found that all covariates were significantly different from LC No Data.

Table 7. Results of a Kruskal-Wallis Posthoc Dunn Test with Holm's correction comparing variable class importance globally, as measured by Per.Inc.MSE, to the variable class Land Cover - No Data.

Variable Class vs. LC No Data	***p*-value**
Built Env. & Urban/Suburban Proxy	< 0.0000
Class of Pop. Place	< 0.0000
Clim./Ecolog./Topo.	< 0.0000
Facilities & Services	< 0.0000
LC Cult. & Managed	< 0.0000
LC Nat. & Semi. Nat. Veg.	< 0.0000
LC Nat. Bare Surfaces	< 0.0000
LU Gen. Class. Var.	0.0003
LU Non-Residential	0.0130
LU Protected	< 0.0000
LU Residential	0.0031
Places & POI	< 0.0000
Pop. Place & Small Poly. Data	< 0.0000
Rivers/Waterbodies/Waterways	< 0.0000
Transportation Network	< 0.0000
Urban/Suburban Extents	< 0.0000

Non-zero Inflated Importances

Plotting the Per.Inc.MSE of only variables included in a given country's final model, for all countries sampled, the distribution by variable class appears as presented in Figure 7. The variability of the mean and median importances of most categories can be seen to increase, relative to the zero-inflated data in Figure 6. Figure 7 presents the distribution of the observed importances of the variable classes only for variables that were included in the final model of a given country's population density. The top five categories for predicting population density by median Per.Inc.MSE becomes Urban/Suburban Extents (10.99%), Clim./Ecolog./Topo. (10.60%), Built Env. & Urban/Suburban Proxies (9.75%), Places & POI (9.50%), and Pop. Places Point & Small Poly Data (6.52%).

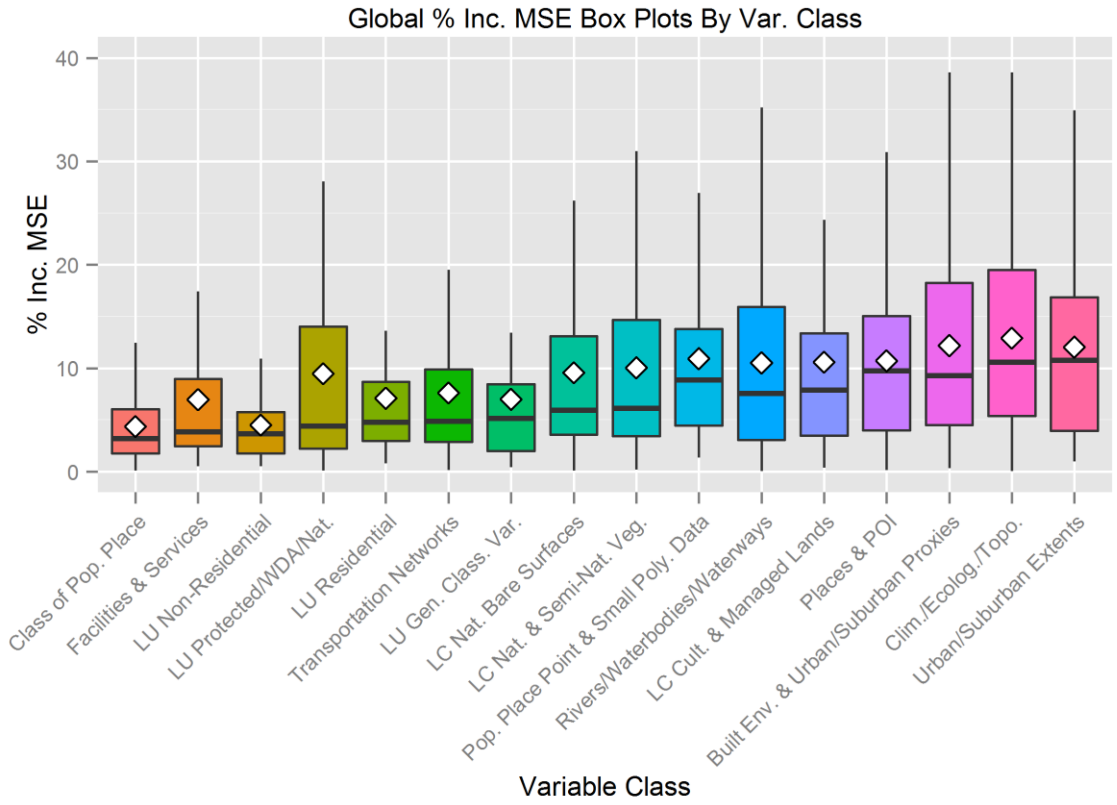


Figure 7. Percent increase in the mean squared error of each covariate class, based upon covariates included in a given country's final model. The mean is represented by a white diamond, the median is represented by the black bar, and the whiskers represent the max and min value within 1.5 * IQR.

When plotting the Per.Inc.MSE for each variable class by region, with the interquartile range (IQR) given by brackets, as done in Figure 8, it can first be noted that many of the variable class IQRs overlap between regions. Further inspection reveals patterns of difference between regions that can be quite distinct. For instance, Facilities & Services variables are observably more important in AFR and SEA as compared to CAC and SAM. Additionally, the amount of variance in the importance of any given variable class, that is the width of the IQR, seems to vary regionally, with CAC and SAM having the least variance across all classes and AFR and SEA having the widest variance across all classes, with few exceptions. More interestingly, it can be seen that all regions tend to exhibit less *between* region variation in the importance of Built Env. & Urban/Suburban variables and for Urban/Suburban Extent variables compared to all other variable classes.

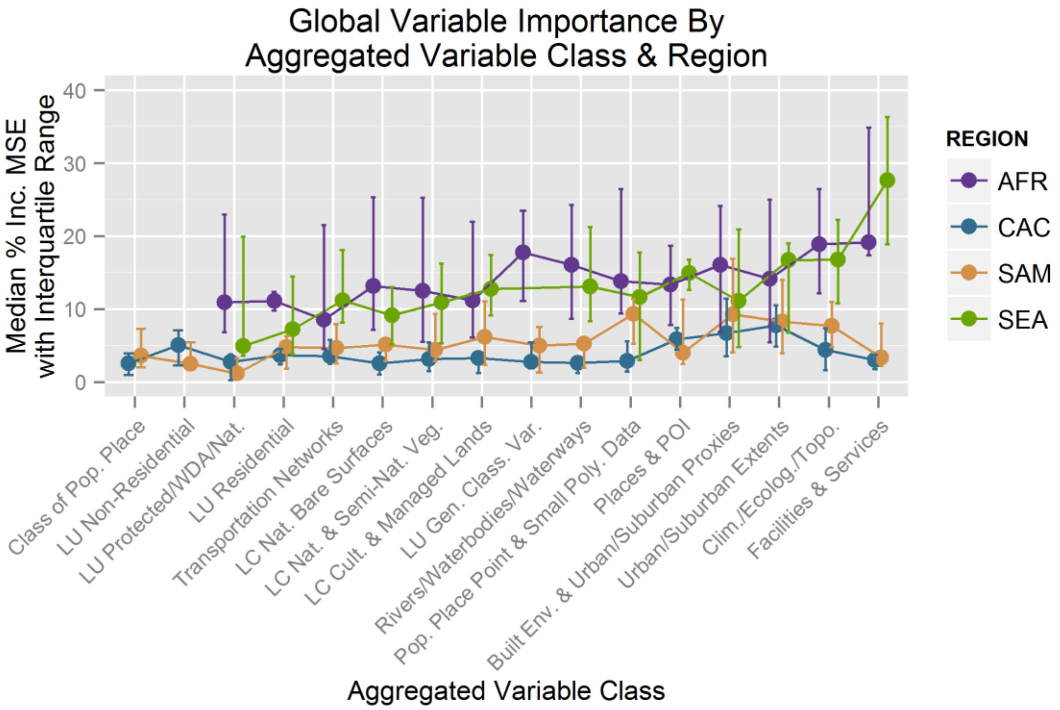


Figure 8. Regional line and dot plot of variable class percent increase in mean squared error (MSE) with the median marked by the dot and the IQR demarcated by brackets.

The importance and variation of importance for each variable class within a region are illustrated in Figure 9. Facilities & Services variables in AFR have the least variation in importance and are relatively strong predictors of population density. Places & POI variables exhibit similar behavior in the SEA region.

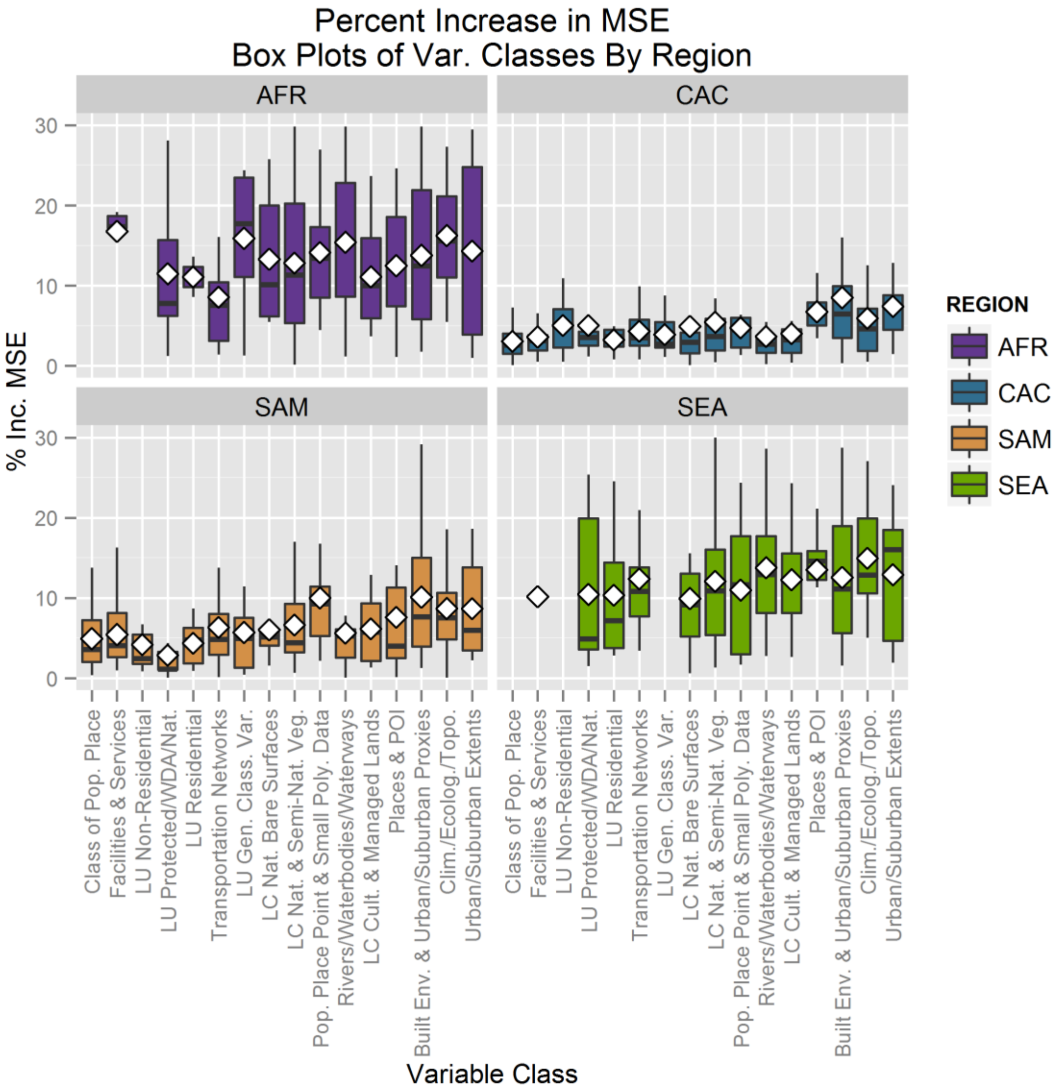


Figure 9. Boxplots of the percent increase in mean squared error (MSE) of each variable category by region. The mean is represented by a white diamond, the median is represented by the black bar, and the whiskers represent the max and min value within 1.5 * IQR.

The Facilities & Services variable class in SEA has such low variation because there were relatively few variables of that class in the country population models, i.e. small sample size (n = 2). CAC has, relative to other regions, little variation in importance across all variable categories, however its strongest predictive variable classes, as judged by the median importance, tend to have greater variation than its weaker predictive variable classes. Note that not all regions contain variables of all classes.

Inter-regional testing for significant differences between countries of a given region and a given variable class were carried out using the Per.Inc.MSE. Significant differences were found between countries in all regions for the variable classes of Clim./Ecolog./Topo., Facilities & Services, and LC Nat. & Semi. Nat. Veg., however, these significant differences either disappeared when correcting for multiple outcomes or were not found when taking into account the total number of covariates in a country's final model by repeating the tests with the weighted importance rank. In fact, no significant differences were detected between countries of a given region and a given variable class when the tests were performed with the weighted importance rank and therefore the results are not shown.

Similarly, within the intra-regional testing using Per.Inc.MSE, some significant differences did not persist when the intra-regional tests were repeated using the weighted importance rank. Given that the weighted importance rank is a more valid measure of covariate class importance when comparing across countries, see Random Forest Considerations in the Discussion, the results of the intra-regional tests using Per.Inc.MSE are not shown.

Importance of Urban Variable Classes

Globally, I compared the resolution of the urban variables, i.e. variables within the classes Urban/Suburban Extents and Built Env. & Urban/Suburban Proxies, against their corresponding Per.Inc.MSE. The results of the global scale test are shown in Table 8. I found no significant differences for AFR ($\chi^2 = 4.718$, d.f. = 4, $p = 0.3174$) and CAC ($\chi^2 = 8.10$, d.f. = 5, $p = 0.1507$) regions therefore no post-hoc tests were performed. After accounting for multiple comparisons, the significant difference found for SEA ($\chi^2 = 9.88$, d.f. = 4, $p = 0.0424$) was not found in pair-wise comparisons and therefore no table was constructed for SEA. The post-hoc test results for SAM are presented in Table 9.

Table 8. Results of global pair-wise post-hoc Dunn test with Holm's correction for multiple outcomes of the percent increase in mean squared error (MSE) of urban covariate data compared to their resolution.

	Corrected Z-value (p-value)				
Resolution	**15 arc sec**	**30m**	**300m**	**500m**	**Other**
30m	3.00 (0.032)	---			
300m	3.55 (0.005)	0.50 (1.00)	---		
500m	4.23 (0.000)	0.68 (1.00)	0.11 (1.00)	---	
Other	1.56 (1.00)	0.94 (1.00)	1.38 (1.00)	1.61 (1.00)	---
Vector	4.82 (0.000)	1.26 (1.00)	0.70 (1.00)	0.68 (1.00)	2.10 (0.391)
Global K-W Test Result: d.f. = 5, Chi-square = 29.37, *p* < 0.0000					

Table 9. Results of pair-wise post-hoc Dunn test with Holm's correction for multiple outcomes of the percent increase in mean squared error (MSE) of urban covariate data compared to their resolution, for sample countries in the SAM region.

	Corrected Z-value (p-value)		
Resolution	**15 arc sec**	**30m**	**500m**
30m	0.97 (0.459)	---	
500m	2.15 (0.124)	1.20 (0.459)	---
Vector	3.64 (0.002)	2.76 (0.028)	1.65 (0.295)
SAM K-W Test Results: d.f. = 3, Chi-square = 15.50, *p* = 0.0014			

Globally, all resolutions of urban variables sampled were significantly different from the 15 arc sec resolution ($p < 0.05$), with the exception of the "Other" resolution which largely consisted of unique and hybrid datasets composed of country specific or hybrid built land cover datasets. No other significant differences between urban variable resolution and Per.Inc.MSE was observed at the global scale of analysis. For the SAM region, the only significant differences observed were between urban variables of “Vector” resolution and 15 arc second (p = 0.0016) and Vector resolution and 30m (p = 0.0282).

Variable Class Importance: Weighted Rank

Variable class importance, globally, as measured by the within-country weighted rank of Per.Inc.MSE, is presented in Figure 10. A weighted rank of zero is the highest importance and takes into account the total number of covariates in a given country’s final model. It can be seen that, relative to the plots of Figure 6 and Figure 7, by accounting for the total number of covariates in a given country’s model the relative importance of the covariate classes shifts. The five most important variable classes, in descending order, for predicting population density by median weighted rank are: Urban/Suburban Extents (0.28), Built Env. & Urban/Suburban Proxies (0.33), Clim./Ecolog./Topo. (0.37), Pop. Place Point & Small Poly. Data (0.42), and Transportation Networks (0.44).

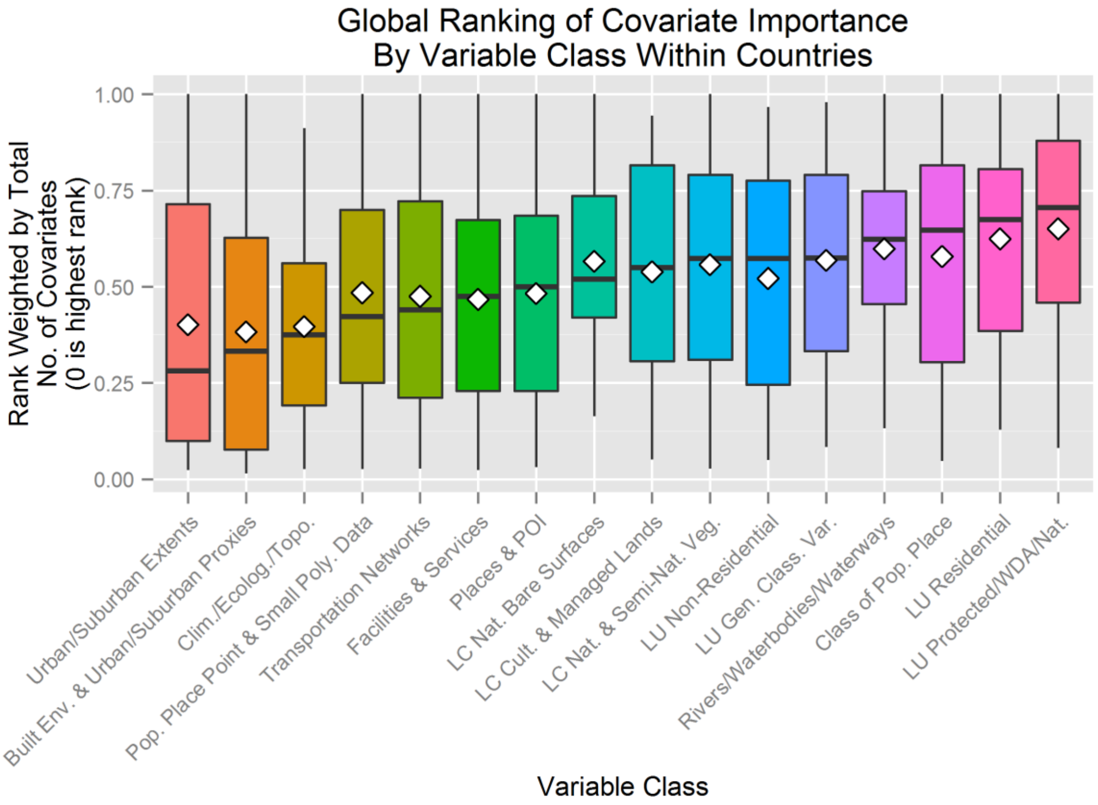


Figure 10. Variable class weighted rank of importance based upon covariates included in a given country's final model. The mean is represented by a white diamond, the median is represented by the black bar, and the whiskers represent the max and min value within 1.5 * IQR.

Accounting for the total number of covariates in a given country's final population model, by converting the importance scores to weighted importance ranks, normalized the variances of each variable category, as shown in Figure 11. When taking the total number of covariates into account, Urban/Suburban Extents and Built Env. & Urban Suburban Proxies rise in importance, as based upon the median weighted rank, across all regions. These two variable classes would appear to be more important and consistently important predictors, based upon the median weighted rank and the variance of their weighted ranks, in the CAC and SAM regions as compared to the AFR and SEA regions.

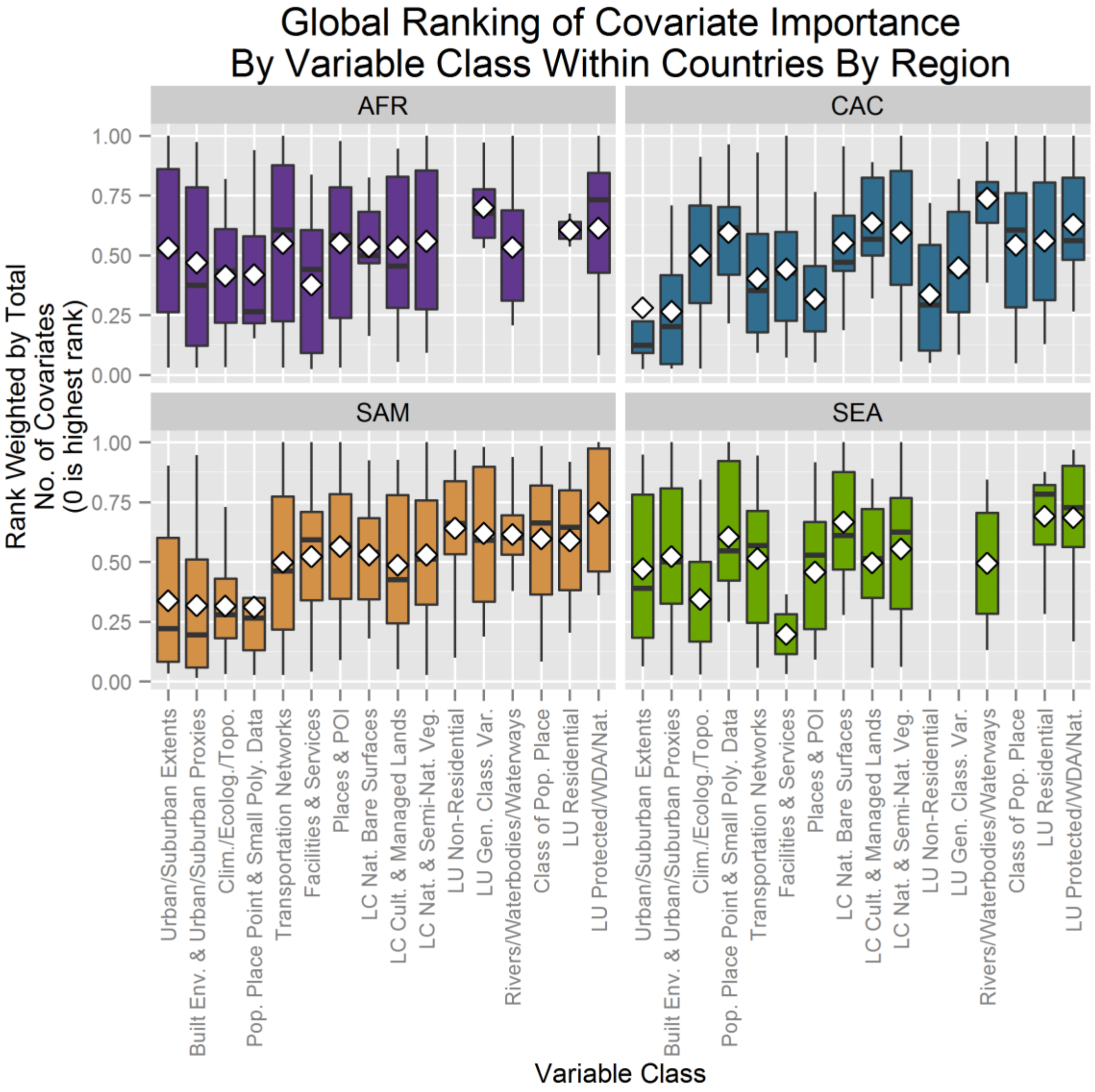


Figure 11. Boxplots of the weighted importance rank of variable classes by region. The mean is represented by a white diamond, the median is represented by the black bar, and the whiskers represent the max and min value within 1.5 * IQR.

Investigating for significant differences between covariate classes globally and intra-regionally, I discovered that significant differences existed globally and within all regions except AFR. Selected comparisons of the top five important covariate classes are presented in Tables 10, 11, and 12, corresponding to the global test, and the CAC and SAM intra-regional tests. The significant differences found in the Kruskal-Wallis test for SEA (χ^2 = 24.42, d.f. = 12, p = 0.0178), were not found after accounting for multiple comparisons in the post-hoc tests and as such no pair-wise table for SEA is presented. None of the top five covariate classes were significantly different from each other. Additionally, globally as well as for every region, Urban/Suburban Extents and Built Env. & Urban/Suburban Proxies were not significantly different from each other.

Table 10. Selected results of pair-wise post-hoc Dunn test with Holm's correction for multiple outcomes of global weighted importance rank of covariate classes.

			Corrected Z-value (p-values)		
Variable Class	**Built Env. & Urban/Suburb Proxies**	**Clim./ Ecolog./ Topo.**	**Pop. Place & Small Poly Data**	**Transportation Network**	**Urban/Suburb Extents**
Class of Pop. Place	5.38 (0.00)	5.15 (0.00)	1.86 (1.00)	2.82 (0.43)	3.76 (0.01)
Clim./Ecolog. /Topo.	0.47 (1.00)	---	1.80 (1.00)	2.31 (1.00)	0.04 (1.00)
Facilities & Services	2.01 (1.00)	1.69 (1.00)	0.31 (1.00)	0.18 (1.00)	1.29 (1.00)
LC Cult. & Managed	3.43 (0.06)	3.16 (0.15)	0.95 (1.00)	1.39 (1.00)	2.53 (0.97)
LC Nat. & Semi. Nat. Veg.	5.48 (0.00)	5.27 (0.00)	1.55 (1.00)	2.56 (0.90)	3.57 (0.03)
LC Nat. Bare Surfaces	3.66 (0.02)	3.42 (0.06)	1.35 (1.00)	1.82 (1.00)	2.84 (0.41)
LU Gen. Class. Var.	3.28 (0.10)	3.05 (0.21)	1.27 (1.00)	1.64 (1.00)	2.62 (0.78)
LU Non-Residential	1.94 (1.00)	1.73 (1.00)	0.46 (1.00)	0.63 (1.00)	1.55 (1.00)
LU Protected	6.05 (0.00)	5.86 (0.00)	2.94 (0.30)	3.94 (0.00)	4.66 (0.00)
LU Residential	3.61 (0.03)	3.42 (0.06)	1.85 (1.00)	2.22 (1.00)	3.05 (0.21)
Places & POI	2.25 (1.00)	1.95 (1.00)	0.02 (1.00)	0.16 (1.00)	1.53 (1.00)
Pop. Place & Small Poly. Data	2.08 (1.00)	1.80 (1.00)	---	0.18 (1.00)	1.46 (1.00)
Rivers/Waterbodies/Waterways	5.78 (0.00)	5.57 (0.00)	2.25 (1.00)	3.29 (0.09)	4.13 (0.00)
Transportation Network	2.65 (0.71)	2.31 (1.00)	0.18 (1.00)	---	1.62 (1.00)
Urban/Suburban Extents	0.38 (1.00)	0.04 (1.00)	1.46 (1.00)	1.62 (1.00)	---
Global K-W Results: d.f. = 15, Chi-square = 106.88, p < 0.0000					

Table 11. Selected results of pair-wise post-hoc Dunn test with Holm’s correction for multiple outcomes of the weighted importance rank of covariate classes for countries located in Central America and the Caribbean (CAC).

			Corrected Z-values (p-values)		
Variable Class	**Built Env. & Urban/Suburb Proxies**	**Clim./ Ecolog./ Topo.**	**Pop. Place & Small Poly Data**	**Transportation Network**	**Urban/Suburb Extents**
Class of Pop. Place	4.24 (0.00)	5.16 (1.00)	0.63 (1.00)	2.17 (1.00)	3.03 (0.24)
Clim./Ecolog. /Topo.	3.72 (0.02)	---	1.10 (1.00)	1.58 (1.00)	2.61 (0.83)
Facilities & Services	2.38 (1.00)	0.80 (1.00)	1.62 (1.00)	0.51 (1.00)	1.76 (1.00)
LC Cult. & Managed	4.13 (0.00)	1.55 (1.00)	0.39 (1.00)	2.64 (0.78)	3.37 (0.08)
LC Nat. & Semi. Nat. Veg.	5.30 (0.00)	1.54 (1.00)	0.04 (1.00)	3.19 (0.14)	3.72 (0.02)
LC Nat. Bare Surfaces	3.02 (0.24)	0.57 (1.00)	0.35 (1.00)	1.59 (1.00)	2.48 (1.00)
LU Gen. Class. Var.	1.99 (1.00)	0.55 (1.00)	1.32 (1.00)	0.49 (1.00)	1.59 (1.00)
LU Non-Residential	0.61 (1.00)	1.44 (1.00)	2.01 (1.00)	0.61 (1.00)	0.46 (1.00)
LU Protected	4.13 (0.00)	1.48 (1.00)	0.30 (1.00)	2.60 (0.85)	3.34 (0.08)
LU Residential	1.96 (1.00)	0.35 (1.00)	0.26 (1.00)	1.01 (1.00)	1.75 (1.00)
Places & POI	0.60 (1.00)	2.05 (1.00)	2.57 (0.92)	0.97 (1.00)	0.40 (1.00)
Pop. Place & Small Poly. Data	3.76 (0.01)	1.10 (1.00)	---	2.22 (1.00)	3.03 (0.24)
Rivers/Waterbodies/Waterways	6.55 (0.00)	3.36 (0.08)	1.50 (1.00)	4.77 (0.00)	4.96 (0.00)
Transportation Network	2.24 (1.00)	1.58 (1.00)	2.22 (1.00)	---	1.51 (1.00)
Urban/Suburban Extents	0.12 (1.00)	2.61 (0.83)	3.03 (0.24)	1.51 (1.00)	---
CAC K-W Test Results: d.f. = 15, Chi-square = 81.28, *p* < 0.0000					

Table 12. Selected results of pair-wise post-hoc Dunn test with Holm's correction for multiple outcomes of the weighted importance rank of covariate classes for countries located in South America (SAM).

			Corrected Z-value (p-values)		
Variable Class	**Built Env. & Urban/Suburb Proxies**	**Clim./ Ecolog./ Topo.**	**Pop. Place & Small Poly Data**	**Transportation Network**	**Urban/Suburb Extents**
Class of Pop. Place	4.84 (0.00)	5.16 (0.00)	3.11 (0.19)	1.91 (1.00)	3.11 (0.19)
Clim./Ecolog./Topo.	0.00 (1.00)	---	0.07 (1.00)	3.07 (0.21)	0.23 (1.00)
Facilities & Services	3.03 (0.24)	3.17 (0.15)	2.15 (1.00)	0.40 (1.00)	2.04 (1.00)
LC Cult. & Managed	1.83 (1.00)	1.87 (1.00)	1.50 (1.00)	0.13 (1.00)	1.34 (1.00)
LC Nat. & Semi. Nat. Veg.	3.52 (0.04)	3.73 (0.02)	2.35 (1.00)	0.59 (1.00)	2.26 (1.00)
LC Nat. Bare Surfaces	2.16 (1.00)	2.20 (1.00)	1.80 (1.00)	0.33 (1.00)	1.66 (1.00)
LU Gen. Class. Var.	3.26 (0.11)	3.34 (0.09)	2.64 (0.78)	1.36 (1.00)	2.55 (1.00)
LU Non-Residential	3.27 (0.00)	3.34 (0.09)	2.70 (0.65)	1.49 (1.00)	2.61 (0.82)
LU Protected	4.18 (1.00)	4.28 (0.00)	3.36 (0.08)	2.32 (1.00)	3.32 (0.09)
LU Residential	2.30 (0.28)	2.34 (1.00)	2.02 (1.00)	0.79 (1.00)	1.90 (1.00)
Places & POI	2.98 (1.00)	3.07 (0.21)	2.31 (1.00)	0.83 (1.00)	2.21 (1.00)
Pop. Place & Small Poly. Data	0.07 (1.00)	0.07 (1.00)	---	1.98 (1.00)	0.24 (1.00)
Rivers/Waterbodies/Waterways	4.02 (0.00)	4.17 (0.00)	2.96 (0.29)	4.77 (1.00)	4.96 (0.34)
Transportation Network	2.90 (0.35)	3.07 (0.21)	1.98 (1.00)	---	1.51 (1.00)
Urban/Suburban Extents	0.23 (1.00)	0.23 (1.00)	0.24 (1.00)	1.51 (1.00)	---
SAM K-W Test Results: d.f. = 15, Chi-square = 69.10, *p* < 0.0000					

DISCUSSION

Variable Class Importance

Given the more valid representation of variable class importance by the weighted importance rank, the discussion of the results from testing first and third hypotheses, from Table 2, will refer to the tests performed with the weighted importance ranks. However, some of the discussion does make reference to the graphical distributions of the Per.Inc.MSE values for the variable classes. These distributions can provide some insight into the underlying data used and how the covariates are interacting within the context of the random forest framework.

As measured by both Per.Inc.MSE and the weighted importance rank, Urban/Suburban Extents, Built Env. & Urban/Suburban Proxies, and Clim./Ecolog./Topo. were consistently seen as the most important predictive variable classes for population density. This result was observed at both the global and intra-regional scales of analysis. Additionally, these three variable classes constituted three of the five variable classes with the highest representation in final population models (Table 6). Table 6 and Figure 10 show that while Urban/Suburban Extents are highly important, if not the most important when looking at the median weighted rank of Per.Inc.MSE, they are less likely to be included (0.5385) in a final population model than Built Env. & Urban/Suburban Proxies (0.5864) which are the second most important by weighted rank of Per.Inc.MSE. More generally, these results indicate that while some variable classes are more likely to be included in a model, they are not necessarily highly important, or strong, predictors of population density in all cases. The opposite is true as well. For instance, Facilities & Services class variables were included in only 38.79 percent of the sampled countries' final models, but in some specific instances they can be highly important predictors of population density. A specific instance is Kenya, where in the final model the distance-to-schools covariate had a Per.Inc.MSE of 34.90 percent and a weighted rank of Per.Inc.MSE of 0.091, ultimately being the second most important predictor of population density for the country. Some of the variance in variable class importances can be explained by the data completeness and quality of a given data set, with the Kenyan school data being an exceptionally complete and accurate dataset. Such data quality characteristics likely explains, in part, the variations seen in regional variable class importances (Figures 8, 9, and 11).

My finding that Built Env. & Urban/Suburban Proxies and Urban/Suburban Extents variable classes were the most important in predicting population density aligns with expectations, especially given that it is estimated that 54 percent of the world's population live in urban areas (U.N. 2014b). Additionally, there are numerous examples in the literature that population or population density and population growth covariates were important in predicting urban area extent (Foresman, Pickett, & Zipperer 1997; López et al. 2001; Chabaeva, Civco, & Prisloe 2004; Herold, Couclelis, & Clarke 2005; Jat, Garg, Khare 2008). This study shows that the relationship, while its exact structure remains unknown, goes in the other direction as well with urban area extent being important in predicting population density. Additionally, transportation and elevation related covariates were found to be of predictive importance for urban land cover, similar to how I found that Transportation Network and Clim./Ecolog./Topo. variable classes were consistently important to predicting population density (Huang, Xie, & Tay 2010; Thapa & Murayama 2011; Linard, Tatem, & Gilbert 2013).

An unexpected finding was how important the Clim./Ecolog./Topo. variable category was in predicting population density, second only to Urban/Suburban Extents and Built Env. & Urban/Suburban Proxies categories. While the category was not broken up for subsequent testing, from examining the covariate importance plots of individual countries, I believe that the majority of this importance is driven by elevation covariates. This also includes elevation derived covariates such as slope.

Additionally, I also showed that the regional definitions used did not display any inter-regional significant differences in the importance of variable classes, which would imply that the definitions were optimal for intra-regional hypothesis testing. It would appear that, based upon regional breakdowns of variable class importance in population density prediction, the AFR and SEA regions and the CAC and SAM regions display similar variable class importances as related to population density (Figure 9).

Importance of Urban Variable Classes

There were few significant differences found between the differing resolutions of the urban variable classes with the primary difference at the global level being whether or not the data was or was not below 15 arc sec in resolution and at the regional level the only significant differences existed in the SAM region between vector resolution data and 15 arc sec and 30m data. These differences may be a result of the original resolution of the primary datasets for each of those classes: 500m MODIS derived urban extents for the Urban/Suburban Extents class and 15 arc sec (~ 463m at the equator) Suomi-VIIRS lights-at-night data for the Built Env. & Urban/Suburban Proxies class (Schneider, Friedl, & Potere 2010; Miller et al. 2012). Or it could be a result of the complex non-linear relationship of these two disproportionately represented datasets, which capture some portion of the variability of population density other variable classes do not.

Alternatively, the observed differences may be a result of the operational definition of “urban” utilized in the construction of any one of the datasets in the Urban/Suburban Extents class. Future work might contribute to this by disentangling the derivation of built- and urban-area definitions as it relates to where people live. Further bias could have been introduced by the very fact that some datasets, such as Suomi-VIIRS lights-at-night data, the 500m MODIS derived urban extents data, and the built land cover derived classifications, are included in every model. Also, the fact that prior to being input into the RF model all covariates are aggregated to the administrative unit resolution may indicate that any effect the original data resolution may have had on the importance is being obfuscated by the resolution of the census data (Stevens et al. 2015).

The current urban datasets utilized lack internal heterogeneity within continuous urban areas which have varied land uses, building structures, and densities. The limited dimensionality of these variable classes will begin to be remedied with the coming release of high-resolution synthetic aperture radar data from sensors such as those aboard the European Space Agency’s Sentinel-1 satellite mission, building footprint data identified through the forthcoming Global Urban Footprint data, and growing land-use data repositories (e.g. Open Street Map, national, regional, local government data, etc.)

that could more appropriately capture the high internal heterogeneity of urban and suburban areas which are composed of varying building heights, varying land uses, and varying patch-like patterns of buildings (Esch et al. 2013). However, even if a 100 percent accurate global building footprint dataset were available tomorrow, the incorporation of accurate ancillary datasets would still need to be included as all buildings are not used for habitation and not all habituated buildings have similar population densities.

Ultimately, I believe that the results of the testing of the potential effect of data resolution on the importance of urban covariates are inconclusive due to the large number of factors that cannot be accounted for.

Issues of Scale

It is important to note that all of these findings are at a specific spatial resolution and modeling scale that may or may not maintain the same forms, structures and relationships at a finer scale as is typically the case with the Modifiable Areal Unit Problem (MAUP) (Openshaw 1984). This may especially hold true for the datasets that currently comprise the Built Env. & Urban/Suburban Proxies and Urban/Suburban Extent variable classes, which are either binary (e.g. urban or non-urban) or give an indication of urban “intensity” (e.g. more intense light measured at night) (Schneider, Friedl, & Potere 2010; Miller et al 2012). However, all covariates are affected to some degree because they are all resampled to 100m prior to being input into the RF model (Stevens et al. 2015).

An additional consideration is the fact that the RF model is determining the relationships between the covariates and population density at the administrative unit level, but is predicting at the smaller pixel (100m) level (Stevens et al. 2015). The relationships at the administrative level may not persist at the finer 100m scale. However, no information is currently available on that other than specific countries for which validation data, of a finer administrative unit scale than which the model was created on, was procured and used to validate the accuracy of the population density predictions.

Referring back to the variance of importances within variable classes and the similar patterns of importance and importance variance between AFR and SEA as well as between SAM and CAC, shown in Figure 8, a partial explanation may lie within the typical number of administrative units used in the regions. Looking at Table 5, it can be seen that AFR and SEA have mean number of administrative units in a country modeled within those regions as 4680.88 and 13746.14 where as in CAC and SAM the mean number of administrative units in a country modeled within those regions as 346.50 and 1020.11. The width of the variances of the importances of the variable categories visually, positively correlates with the increasing average number of administrative units used in modeling the countries of those regions.

This makes sense due to the scale effect of the MAUP, which generally states that as you decrease the number of areal units there is a decrease in the variability of the observations corresponding to the areal units (Openshaw 1984). The potential of the ASR (Table 5) in having some effect on this variability is less clear, but likely has an effect relative to the concept of the MAUP zonation effect (Openshaw 1984). So in addition to accounting for data completeness and quality when trying to account for the variability of covariate importances, the number of administrative units being used in the modeling process and the ASR of those units should be accounted for in some manner. Further explicit investigation into the effect the number of admin units used in modeling and the effect of the ASR should be conducted.

Random Forest Considerations

There are inferential limits to using the RF model to identify/approximate the structure and nature of variable class relationships to population density. As Breiman (2001, p20) stated, “A forest of trees is impenetrable as far as simple interpretations of its mechanism go.” Unlike multiple linear regressions or a singular classification and regression tree (CART) where coefficients and confidence intervals can be quantified or decision paths can be traced from input observation to CART predictions, the numerous (typically 500 or more) trees in a RF preclude the tracing of the regression of input to prediction (Breiman 2001). Furthermore, the strength of a RF to capture highly non-linear relationships of covariates and their complex interactions, which allows for more accurate predictions, does not lend itself to simple interpretations of the underlying mechanisms of the modeled phenomenon (Breiman 2001). Variable importance within a RF is similarly complex due to those same non-linear relationships and intricate interactions amongst covariates (Liaw and Wiener 2002). This results in the effect of a covariate’s importance in a RF model being highly conditional on all the other covariates present, with similar results not being guaranteed in other models, even for the same country.

The strengths of a RF in a population modeling context far outweigh its limitations if the priority is to accurately predict population density rather than ascertain in-depth understanding of the mechanisms between population density and the covariates used to predict it. Given the numerous potential variables used to model population, with many containing only a small amount of additional information, a RF provides improved accuracy where a single tree classifier would only provide accuracy slightly better than random (Breiman 2001). Through studies like this, where numerous random forests modeling population density, better ways of identifying and addressing inherent bias in predictions can be attained. In addition to being robust to noise and small datasets, RFs do not over fit the data due to the Law of Large Numbers (Breiman 2001). This is characteristic is particularly useful for countries where only coarse census data is available, i.e. relatively few administrative units with a large ASR. The strengths, and limits, of using an RF model in a population modeling context being stated, I can still come to global and regional conclusions regarding the general patterns of variable class importance for modeling population density at 100m resolution for countries even if I cannot come to conclusions pertaining to the underlying mechanisms and interactions driving these importances.

CONCLUSIONS

This study has quantified what has often been taken as common knowledge: that urban areas are the best predictors of where to find high population densities. I have found that Built Env. & Urban/Suburban Proxies, Urban/Suburban Extents, and Clim./Ecolog./Topo. variable classes are the most important to predicting population density, both globally and regionally. There are some slight regional variations in the patterns of variable class importance amongst the variable classes found to be of middling or low predictive importance, but overall there is little significant interregional difference in these patterns. However, the exact mechanism and structure of the underlying relationship(s) between these variable classes and population density are not discernable within the RF method. Additionally, these patterns of variable class importance are for a specific spatial resolution and modeling scale which could or could not maintain their form and relationship at a finer scale.

Next steps in further investigating these variables in relation to population density could involve utilizing a different modeling framework which would allow for more inferential power as to the structure and nature of the relationships between these variables and population density. Additionally, focusing study on specific variable classes, such as the urban/suburban related variable classes, by sourcing novel and forthcoming datasets that help illuminate the heterogeneity of these areas, both internally and across different countries and regions, could increase the predictive ability of a population model regardless of the framework.

Based upon the results of this study, priorities for improving the accuracy of population maps would be sourcing high resolution settlement datasets, encouraging development and release of more detailed census data, and investigate the availability or development of important predictor covariate datasets, on a country-by-country basis, that currently are not performing as well as the regional average. Overall, a more in-depth characterization of population density and predictive covariates are needed. Another investigation that is warranted is to determine if and what covariates of low predictive importance can be consistently dropped from current modeling in an effort to increase the end-user utility of these datasets.

REFERENCES

Alegana, V. A., P. M. Atkinson, C. Pezzulo, A. Sorichetta, D. Weiss, T. Bird, E. Erbach-Schoenberg & A. J. Tatem. 2015. Fine resolution mapping of population age-structures for health and development application. *Journal of the Royal Society Interface,* 12.

Agresti, G. 2007. An Introduction to Categorical Data Analysis (2nd ed.). Hoboken, NJ, USA: John Wiley & Sons, Inc.

ArcGIS Desktop Release 10.2. Environmental Systems Research Institute, Redlands, CA.

Balk, D., U. Deichmann, G. Yetman, F. Pozzi, S. I. Hay & A. Nelson. 2006 Determining global population distribution: Methods, applications, and data. *Advanced Parasitology,* 62**,** 119-156.

Balk, D. & G. Yetman. 2004. The global distribution of population: Evaluating the gains in resolution refinement.

Bhaduri, B., E. Bright, P. Coleman & M. L. Urban. 2007. LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal,* 69**,** 103-117.

Bharti, N., X. Lu, L. Bengtsson, E. Wetter & A. J. Tatem. 2015. Remotely sensing population during a crisis by overlaying two data sources. *International Health,* 7**,** 90-98.

Breiman, L. 1996. Bagging predictors. Machine Learning 24(2): 123–140.

---. 2001. Random Forests. *Machine Learning,* 45**,** 5-32.

Briem, G.J., Benediktsson, J.A., & J.R. Sveinsson. 2002. Multiple classifiers applied to multisource remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing* 40(10): 2291–2299.

Buhaug, H. & H. Urdal. 2013. An urbanization bomb? Population growth and social disorder in cities. *Global Environmental Change,* 23**,** 1-10.

Chabaeva, A. A., D. L. Civco, & S. Prisloe. 2004. Development of a population density regression model to calculate imperviousness. In: *ASPRS Annual Conference Preceedings*, Denver, CO , USA.

Chan, J.C.-W., Paelinckx, D., 2008. Evaluation of Random Forest and Adaboost treebased ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment 112(6): 2999–3011.

Cheriyadat, A., E. Bright, D. Potere & B. Bhaduri. 2007. Mapping of settlements in high-resolution satellite imagery using high performance computing. *GeoJournal,* 69**,** 119-129.

Chongsuvivatwong, V., K. H. Phua, M. T. Yap, N. S. Pocock, J. H. Hashim, R. Chhem, S. A. Wilopo & A. D. Lopez. 2011. Health and health-care systems in southeast Asia: Diveristy and transitions. *The Lancet,* 377**,** 429-437.

CIESIN. 2011. Global Rural Urban Mapping Project (GRUMP). sedac.ciesin.columbia.edu: Center for International Earth Science Information Network (CIESIN).

Cohen, B. 2006. Urbanization in developing countries: Current trends, future projections, and key challenges for sustainability. *Technology in Society,* 28**,** 63-80.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In "Multiple Classifier Systems". Vol. 1857 of the series *Lecture Notes in Computer Science*: 1-15

Dobson, J. E., E. Bright, P. Coleman, R. C. Durfee & B. A. Worley. 2000. LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering & Remote Sensing,* 66**,** 849-857.

Doxsey-Whitfield, E., K. MacManus, S. B. Adamo, L. Pistoles, J. Squires, O. Borkovska, S. R. Baptista. 2015. Taking adantage 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.

Dunn, O.J. 1964. Multiple comparisons using rank sums. *Technometrics* 6: 241-252.

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

Elvidge, C. D., M. L. Imhoff, K. E. Baugh, V. R. Hobson, I. Nelson, J. Safran, J. B. Dietz & B. T. Tuttle. 2001. Night-time lights of the world: 1994-1995. *Photogrammetry & Remote Sensing,* 56**,** 81-99.

Esch, T., M. Marconcini, A. Felbier, A. Roth, W. Heldens, M. Huber, M. Schwinger, H. Taubenböck, A. Müller, & S. Dech. 2013. Urban footprint processor - Fully automated processing chain generating settlement mask from global data of the TanDEM-X mission. *IEEE Geoscience and Remote Sensing Letters* 10(6). doi: 10.1109/LGRS.2013.2272953.

Foresman, T. W., S. T. A. Pickett, & W. C. Zipperer. 1997. Methods for spatial and temporal land use and land cover assessment for urban ecosystems and application in the greater Baltimore-Cheasapeake region. *Urban Ecosystmes* 1(4): 201-216.

Gaughan, A. E., F. R. Stevens, C. Linard, P. Jia & A. J. Tatem. 2013. High resolution population distribution maps for southeast Asia in 2010 and 2015. *PLoS One,* 8**,** e55882.

Herold, M., H. Couclelis, K. C. Clarke. 2003. The role of spatial metrics in the analysis and modeling of urban land use change. *Computers, Environment, and Urban Systems* 29: 369-399.

Herold, M., N. C. Goldstein, & K. C. Clarke. 2003. The spatiotemporal form of urban growth: Measurement, analysis, and modeling. *Remote Sensing of Environment* 86: 286-302.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandanavian Journal of Statistics* 6(2): 65-70.

Huang, B., C. Xie, & R. Tay. 2010. Support vector machines for urban growth modeling. *Geoinformatica* 14: 83-99.

Jat, M. K., P. K. Garg, & D. Khare. 2008. Monitoring and modelling of urban sprawl using remote sensing and GIS techniques. *Int. Journal of Applied Earth Observation and Geoinformation* 10: 26-43.

Kruskal, W. H. & W. A. Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47: 583-621.

Langford, M., D. J. Maguire, & D. J. Unwin. 1991. The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In: *Handling Geographic Information,* 55-77. Essex, U.K.:Longman Scientific and Technical.

Linard, C., M. Gilbert, R. W. Snow, A. M. Noor & A. J. Tatem. 2010. Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS One*, 7**,** e31743.

Linard, C., A. J. Tatem, & M. Gilbert. 2013. Modelling spatial patterns of urban growth in Africa. *Applied Geogrpahy* 44: 23-32.

López, E. G. Bocco, M. Mendoza, E. Duhau. 2001. Predicting land-cover and land-use change in the urban fringe: A case in Morelia city, Mexico. *Landscape and Urban Planning* 55: 271-285.

Madlener, R. & Y. Sunak. 2011. Impacts of urbanization on urban structures and energy demand: What can we learn for urban energy lanning and urbanization management? *Sustainable Cities and Society,* 1**,** 45-53.

Masters, W. A., A. A. Djurfeldt, C. De Haan, P. Hazell, T. Jayne, M. Jirström & T. Reardon. 2013. Urbanization and farm size in Asia and Africa: Implications for food security and agricultural reSEArch. *Global Food Security,* 2**,** 156-165.

McGranahan, G., D. Balk & B. Anderson. 2007. The rising tide: Assessing the risks of climate change and human settlements in low elevation coastal zones. *Environment and Urbanization,* 19**,** 17-37.

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

Miller, S. D., S. P. Mills, C. D. Elvidge, D. T. Lindsey, T. F. Lee, & J. D. Hawkins. 2012. Suomi satellite brings to light a unique frontier of nighttime environmental sensing capabilities. *PNAS* 109(39): 15706-15711.

Meyer, W. B. & B. L. Turner. 1992. Human Population Growth and Land-Use/Cover Change. *Annual Review of Ecology and Systematics,* 23**,** 39-61.

Neuwirth, E. 2014. RColorBrewer: ColorBrewer Palletes. R package version 1.1-2. http://CRAN.R-project.org/package=RColorBrewer

Openshaw, S. 1984. The Modifiable Areal Unit Problem. In *Concepts and Techniques in Modern Geography, No. 38*. Geobooks: Norwich, England.

Pohlert, T. 2016. Calculate pairwise multiple comparisons of mean rank sums. R package version 4.1. http://CRAN.R-project.org/package=PMCMR

Pozzi, F. & C. Small. 2005. Analysis of urban land cover and population density in the United States. *Photogrammetric Engineering & Remote Sensing,* 71**,** 719-726.

Ramankutty, N., J. A. Foley & N. J. Olejniczak. 2002. People on the land: Changes in global population and croplands during the 20th century. *Ambio,* 31**,** 251-257.

RCore Team. 2015. R: A Language and Environment for Statistical Computing Vienna, Austria.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo, & J. P. Rigol-Sanchez. 2012. An assessment of the effectiveness of a random forest classifier for landcover detection. *ISPRS Journal of Photogrammetry and Remote Sensing* 67: 93-104.

Rosner, B. 2011. *Multisample Inference.* In Fundamentals of Biostatistics, 7th, ed. M. Taylor, 516-576. Boston, MA: Brooks/Cole

Schneider, A., M. A. Friedl, & D. Potere. 2010. Mapping global urban areas using MODIS 500-m data: New methods based on 'urban ecoregions'. *Remote Sensing of Environment* 114: 1733-1746.

Small, C. & R. J. Nicholls. 2003. A global analysis of human settlement in coastal zones. *Journal of Coastal Research,* 19**,** 584-599.

Sorichetta, A., G. M. Hornby, F. R. Stevens, A. E. Gaughan, C. Linard & A. J. Tatem. 2015. High-resolution gridded population distribution datasets of Latin America in 2010, 2015, and 2020. *Scientific Data* 2: 150045. doi:10.1038/sdata.2015.45

Stephenson, J., K. Newman & S. Mayhew. 2010. Population dynamics and climate change: What are the links? *Journal of Public Health,* 32**,** 150-156.

Stevens, F. R., A. E. Gaughan, C. Linard & A. J. Tatem. 2015. Disaggregating census data for population mapping using Random Forests with remotely-sensed and ancillary data. *PLoS One,* 10**,** e0107042.

Sverdlik, A. 2011. Ill-health and poverty: a literature review on health in informal settlements. *Environment and Urbanization,* 23**,** 123-155.

Tatem, A. J., J. Campbell, M. Guerra-Arias, L. de Bernis, A. Moran & Z. Matthews. 2014. Mapping for maternal and newborn health: the distributions of women of childbearing age, pregnancies and births. *International Journal of Health Geographics,* 13.

Thapa, R. B., & Y. Murayama. 2011. Urban growth modeling of Kathmandu metropolitan region, Nepal. *Computers, Environment, and Urban Systems* 35: 25-34.

Tobler, W., U. Diechmann, J. Gottsegen, K. Maloy. 1997. World population in a grid of spherical quadrilaterals. *International Journal of Population Geographics* 3: 203-225.

U.N. 2014a. World Population Prospects: The 2014 Revision. In *World Population Prospects*. Washington, D.C.: United Nations.

---. 2014b. World Urbanization Prospects: The 2014 Revision. In *World Urbanization Prospects*. United Nations, Dept. of Economic and Social Affairs, Population Division.

---. 2015. World Population Prospects: The 2015 Revision - Key Findings. In *World Population Prospects*. United Nations.

UNEP. 2004. UNEP/GRID - Souix Falls Clearinghouse. www.na.unep.net/siouxfalls/datasets/datalist.php : United Nations Environment Programme.

UNFPA. 2014. A Universal Pathway. A Woman's Right to Health. In *The State of the World's Midwifery*. United Nations.

Wickham, H. 2009. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

Wickham, H. 2015. tidyr: Easily Tidy Data with `spread()` and `gather()` Functions. R package version 0.3.1. http://CRAN.R-project.org/package=tidyr

Wickham, H. and R. Francois. 2015. dplyr: A Grammar of Data Manipulation. R package version 0.4.3. http://CRAN.R-project.org/package=dplyr

W.H.O. 2014. World Malaria Report. World Health Organization.

WorldPop. 2014. *Ebola*. www.worldpop.org: WorldPop. Last accessed 14 March 2016.

---. 2015a. *What is WorldPop?* www.worldpop.org: WorldPop. Last accessed 14 March 2016.

---. 2015b. *Myanmar flooding*. www.worldpop.org: WorldPop. Last accessed 14 March 2016.

---. 2015c. *Nepal earthquake*. www.worldpop.org: WorldPop. Last accessed 14 March 2016.

Wright, J. K. 1936. A method of mapping densities of population. *The Geographical Review* 26:103-110.

CURRICULUM VITAE

Jeremiah J. Nieves

jeremiah.j.nieves@outlook.com | +1 (502) 640-4507
ResearchGate: researchgate.net/profile/Jeremiah_Nieves
LinkedIn: linkedin.com/in/jeremiahnieves/

Recent Experience

Lecturer, Spatial Statistics – January 2016 to Present
Dept. of Geography & Geosciences, University of Louisville – louisville.edu
Louisville, Kentucky, U.S.A.

Responsibilities:

- Creation of training materials to teach complex mathematical techniques and how to communicate statistical findings to both a specialized and lay-audience
- Leading and teaching statistical methods to a group of 30 students of diverse background and skills
- Supervising and mentoring an undergraduate teaching assistant

Researcher – August 2014 to Present
WorldPop Project – worldpop.org
Based in Southampton, United Kingdom

Responsibilities:

- Conducting research on modeling population and demographics on a global scale using machine learning methods, Bayesian statistical methods, and object-based image analysis methods
- Programming statistical and spatial models in R Statistical Software, Python, and ArcGIS and the arcpy library and integrating scripts to work in a cloud-computing environment (i.e. Google Compute Engine and Microsoft Azure)
- Data mining, data gathering, and data appraisal from open and closed sources
- Collaborating internationally with other researchers, governments, and non-governmental organizations
- Presenting and publishing of research conducted through peer-reviewed journals, academic conferences, and before lay-audiences

Researcher – August 2014 to Present
Dept. Geography & Geosciences, University of Louisville – louisville.edu
Louisville, Kentucky, U.S.A.

Responsibilities:

- Data mining open and closed sources for transportation modeling data
- Planning and executing network-based transportation modeling research design for transportation modeling – the rail and bus public transit system for the entire state of New Jersey for each year between 2005 and 2011
- Modeling travel times for private vehicle networks and multi-modal public transportation networks and managing data for joining to public health data
- Proper handling and management of protected health information

Education

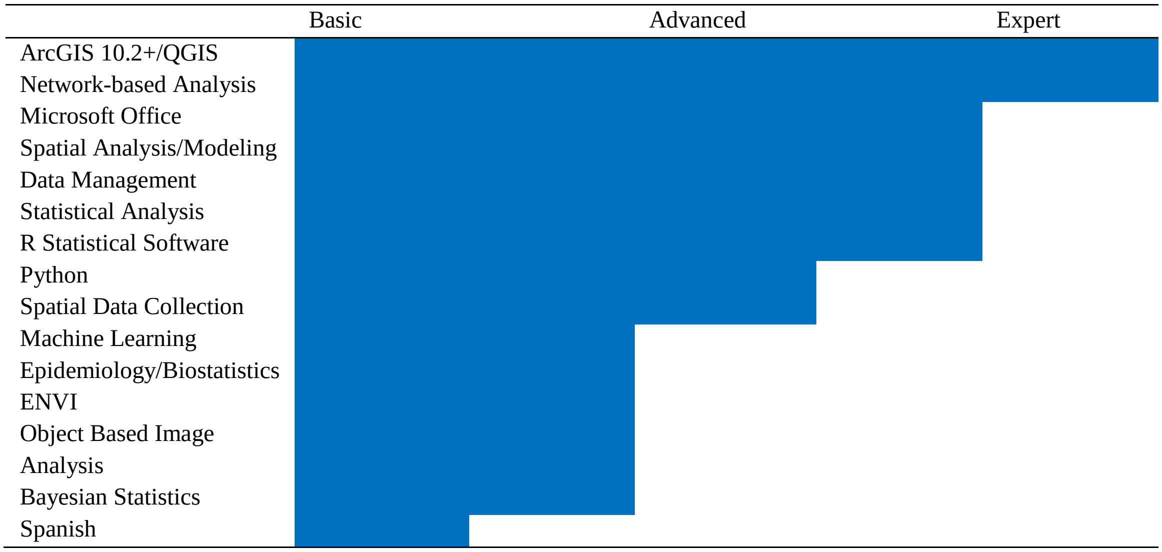
M.S. Applied Geography – August 2014 to Present
University of Louisville – Louisville, Kentucky, U.S.A.
GPA – 4.0

Masters Public Health conc. Biostatistics – August 2013 to May 2014 (Transfer)
University of Kentucky – Lexington, Kentucky, U.S.A.
GPA – 4.0 (no degree awarded; transfer)

B.S. Applied Geography conc. GIS – August 2009 – May 2013
University of Louisville – Louisville, Kentucky, U.S.A.
GPA – 3.757

- Trustee's Scholarship – 2009 to 2013
- 1st Place Geography Poster, 98th Kentucky Academy of Sciences Annual Meeting – 2012
- Dean's Scholar List – 2010 to 2013
- 2nd Place Senior Thesis – 2013

Skills



Publications

Nieves, J. J. and C. A. Day. 2014. Microclimatic and pedologocal variables in rock shelters containing *S. albopilosa*, Red River Gorge, Kentucky. *Environment, Ecology, & Management*, Vol. 2014.

Nieves, J. J. 2015. Combining multi-modal network models with kernel density methods to measure the relative spatial accessibility of pediatric primary care services in Jefferson County, Kentucky. *International Journal of Applied Geospatial Research*.

Gaughan, A. E., F. R. Stevens, Z. Huang, **J. J. Nieves**, A. Sorichetta, S. Lai, X. Ye, C. Linard, G. M. Hornby, S. I. Hay, H. Yu, & A. J. Tatem. 2016. Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Scientific Data* 3. doi: 10.1038/sdata.2016.5

Posters and Presentations

“The Value of Urban Extents in Global Population Mapping”

- 2015 AAG Annual Conference – Chicago, Illinois, U.S.A.

“WorldPop Global Populations and Urban Extents Using Google Compute Engine”

- 2015 AAG Annual Conference – Chicago, Illinois, U.S.A.
- 2015 Kentucky EPSCoR Conference – Lexington, Kentucky, U.S.A.

“Differences in Spatial Access to Pediatric Primary Care Services Among Impoverished Urban and Suburban Communities, Jefferson County, Kentucky”

- 2013 ACC Meeting of the Minds Conference – Winston-Salem, North Carolina, U.S.A.
- 2013 TFISE Conference – Lexington, Kentucky, U.S.A.

“Microclimatic and Pedological Variables in Rock Shelters Containing *S. albopilosa*, Red River Gorge, Kentucky”

- 2013 12th Annual Kentucky Posters at the Capitol – Frankfort Kentucky, U.S.A.
- 2012 University of Louisville Summer Research Opportunity Program – Louisville, Kentucky, U.S.A.
- 2012 98th Annual Kentucky Academy of Sciences Meeting – Richmond, Kentucky, U.S.A.

Membership

- **American Radio Relay League – 2014 to Present**
- **University of Kentucky College of Public Health Research Committee – 2014**
- **Kentucky Assoc. of Mapping Professionals – 2013 to Present**
- **University of Kentucky Public Health Association – 2013 to 2014**
- **Assoc. of American Geographers – 2012 to Present**
- **University of Louisville Geography Club – 2012 to 2013**
- **University of Louisville Student Art League – 2010 to 2011**
    - President, 2011
    - Internal Communication Officer, 2010

References

References available upon request.