Population Density Modeling in Peru: Model Structure Analysis

W. Anderson[1,3*], S. Guikema[1], B. Zaitchik[2], W. Pan[4]

[1] {Department of Geography and Environmental Engineering, The Johns Hopkins University, Baltimore, MD, USA}
[2] {Department of Earth and Planetary Sciences, The Johns Hopkins University, Baltimore, MD, USA}
[3] {Risk Management Solutions, Hoboken, NJ, USA}
[4] {Nicholas School of Environment and Duke Global Health Institute, Duke University}
*Corresponding author email: Weston.B.Anderson@gmail.com

Abstract

Obtaining accurate intra-censal estimates of population is essential for policy and health planning, but is often difficult in countries with limited data. In lieu of available census or other population data, environmental data can be used in statistical models to produce dynamic population estimates. In this study we compare the predictive accuracy of five model structures, including parametric and non-parametric models, to identify statistical modeling structures that most effectively incorporate ancillary data to estimate population density. Environmental covariates include land surface temperature, NDVI and density of rivers, roads or permanent water. Results demonstrate that a regression-based approach is preferred when previous population information is a covariate, but a non-parametric tree-based model provides more accurate estimates when population is used to train the model, but not as a covariate. This latter approach is important for regions with incomplete census data and has implications for economic, health and development policies.

1.0 Introduction:

Estimates of the distribution and growth of human population are invaluable. They are used as input to research-focused and operational applications, including emergency response, infectious disease early warning systems, resource allocation projections and food security analysis, to list only a few examples. However, obtaining reliable population estimates at the spatial and temporal resolutions required for many of these applications is a significant challenge. Census data, the primary source of population size, is limited in temporal frequency and is often incomplete or unreliable - particularly in less-developed countries - which causes considerable problems for policy planning and decision makers. For this reason, models that can refine existing estimates of human populations or that can estimate and project populations in areas that lack population data altogether are of considerable importance.

To date, most approaches to intercensal estimates and postcensal projections can be categorized into one of five methods: cohort component, microsimulation, mathematical formulation, aerial interpolation or statistical modeling. The choice of these methods has often been motivated by both the type of data available (direct vs. indirect data) and the type of estimate requested (total population, subnational, population characteristics, etc.). Although these approaches are not mutually exclusive, they do provide a useful framework for discussion.

The cohort-component method is the most common approach for creating intercensal and postcensal estimates, particularly in developed countries where baseline census data are regularly collected and more information on population growth components are available. This method applies the basic demographic estimating equation ($P_t = P_0 + B - D + [I - E]$), which consists of

estimates of the baseline population (P_0), and (age-specific) births (B), deaths (D) and net migration (I – E) between the baseline and target year to provide population estimates requested. It is flexible to the extent that subgroups of the population can be estimated and projected – i.e., by age and sex, ethnicity, geographic subdivision, etc. When additional characteristics are used to define the cohort, the model is referred to as a multi-state model because individuals being projected forward may enter different states (e.g., married vs. unmarried). These state-specific birth, death and migration rates are obtained in a variety of ways, including civil registrars, sample surveys, prior census data, UN model life tables and expert opinion. The cohort component method of population estimation may capture emerging demographic trends that models based on aggregate measures of population would miss.

When a large number of states are requested for a population estimate or projection, it is often preferable to use microsimulation as opposed to a multi-state cohort component model. Microsimulation produces population estimates by modeling specific individuals and life events of those individuals (O'Neill et al., 2001). Because the method is so computationally intensive, the simulation is often limited to a sample of the population, which is then scaled to match the total population. Microsimulation is impractical when only a limited number of states are desired, but may be preferable for small-scale state-intensive problems.

Mathematical formulation seeks not to model specific individuals or groups of individuals, but rather to define patterns of aggregate population explicitly using mathematical functions, which can be calibrated for a given area, but that more generally describe observed spatial patterns of population density. The most common type of type of mathematical formulation is an estimate

of population growth based on extrapolation of recent census data. The approach has a transparency of formulation that makes it appealing for many applications. One example of an operational extrapolation-based population model is the Gridded Population of the World: Future Estimates 2015 (GPW2015), which produces gridded population estimates globally at 2.5 arc-minute quadrilateral grid resolution (Balk et al., 2005). The product uses a simple geometric extrapolation from the most recent census data (1990 or 2000) on the smallest available sub-national grid. The authors acknowledge that geometric extrapolation is "not typically employed for longer-term projections because it lacks information useful for the longer-term adjustments to population composition and dynamics" (Balk et al., 2005). The GPW2015 was produced as the first spatially distributed global population projection to provide decision makers and researchers information on which they may base their policies and analysis.

Aerial interpolation differs from extrapolation models in that it is used to transform data from one set of spatial units into another, rather than to project trends. In the context of population modeling this entails distributing administrative level census data across a finer scale grid to produce a detailed population surface. The most commonly used technique for producing heterogeneous population density surfaces from homogeneous zones is dasymetric mapping, which uses ancillary information to divide each zone of the source data into subzones (Eicher and Brewer, 2001). Each subzone is assigned a population density such that the sum of population over all subzones equals the population of the original source zone (Langford et al., 1991). The LandScan Global Population Project employs the dasymetric mapping method to disaggregate census population measurements from administrative tracts to 1 km resolution (Dobson et al., 2000).

An interesting feature of LandScan is that the model incorporates a variety of remotely sensed data including information about nighttime lights, uninhabited areas, density of roads, slope and land cover. This use of remotely sensed data from aircraft or satellite has a fairly long tradition in population modeling. Early uses of remote sensing for population estimation were logical extensions of aerial photography, which has been used to count dwellings since the mid 1950s in areas without reliable population information (Boudot, 1993; Puissant, 2010). Following from this, high-resolution remote sensing was used to disaggregate population counts in urban spaces under the assumption that areas with similar land cover will have similar population densities. In recent years, remote sensing has become a prominent source of environmental information, including land use and transportation patterns, which can provide valuable input to population models.

In LandScan, remotely sensed data are used to inform spatial disaggregation of existing population estimates. But remote sensing is perhaps even more valuable for statistical population models in which remotely sensed data are used to estimate population densities as opposed to disaggregating them. The most common way that this is done is to relate the remotely sensed data to land use and to include that information in a regression-based model that is identified and trained using one dataset and evaluated using a separate dataset from a culturally and demographically similar area (Harvey, 2002; Lo, 2003). While remotely sensed data are often used to derive social or economic information relevant to population density, satellite observations may also be included directly in a population model, as demonstrated by Liu and Clarke (2002), who used high resolution satellite-derived reflectance and landscape texture

117 information to estimate population distribution within a single city. The inclusion of these
118 remotely detected data-- either directly or indirectly-- allows modelers to draw insight into the
119 underlying drivers of local population processes. This principle of drawing insight based on
120 model structure differs markedly from the methodology underlying models based on expert
121 knowledge, in which model structure is based on relationships assumed a priori.
122
123 Unlike aerial interpolation and mathematical formulation, statistical modeling does not
124 necessarily use population as a direct input for defining a population surface. Statistical modeling
125 instead focuses on deriving the relationships between population density and external variables
126 by using population data to train a model (Wu et al., 2005). Importantly, the frequency with
127 which updated estimates are produced by a statistical model is constrained by the availability of
128 the covariate data and not by the availability of response variable data. This is a particular
129 advantage for population modeling, for which the only available population information is often
130 the decadal-scale census. By decoupling the temporal resolution of population density estimates
131 from census frequency, modelers can capture the changing dynamics of local populations at a
132 much finer time scale. The ability to move beyond disaggregating static population counts
133 towards predicting population density has spatial benefits as well, allowing modelers to produce
134 population estimates for regions that lack prior population data altogether.
135
136 Previous work has demonstrated the benefit of including ancillary data in statistical population
137 modeling (Sanderson, 1998; Harvey, 2002; Lo, 2003;). Building on that work, and with a
138 particular focus on model structure, we compare the predictive accuracy of five different
139 statistical modeling techniques—including both parametric and non-parametric methods—in an

effort to identify and understand alternative model structures for population prediction in data-limited regions. Each model in the study uses a range of covariates to predict population density at the district level in Peru. In the following sections we will describe data sources and the required data processing (Section 2), detail the structure of the models included (Section 3), present and discuss results (Section 4), and offer general conclusions (Section 5). The analysis presented in this paper reveals that the presence or absence of previous census data transforms the problem of statistical population modeling from one that benefits most from a geometric extrapolation of previous population data to one that benefits most from a non-parametric model and a wide range of covariates. These results are relevant for regions with limited reliable census data, particularly those experiencing rapid population redistribution towards frontier zones, as is the case in many parts of Peru.

2. Data

2.1 Scope of the Study

Peru is divided administratively into regions then provinces followed by districts. The variables used, discussed below, were calculated annually at the district level for five regions (Ayacucho, Cusco, Madre de Dios, Arequipa and Apurimac). Combined, these regions contain 42 provinces, 417 districts and span a reasonable cross section of Peruvian land cover (see Fig. 1). The country exhibits a broad range of climatic variability with land cover including rainforest, mountains and coastal areas. The five regions included in the study were chosen to be representative of Peruvian topography and to avoid regions that were redistricted during the study period. Our application of population modeling to Peru is motivated in large part by the need for accurate population estimates and projections to inform malaria early warning systems and risk assessment currently

being developed for the Peruvian Amazon frontier. Accurate population estimates are critical for understanding and predicting malaria in such frontier zones (de Castro et al., 2007). However, the contributions are more general as will be discussed below.

2.2 Response variable:

Population density in 2007 was used as the response variable and was calculated using the 2007 census data downloaded from the Peruvian Institute of Statistics and Information (El Instituto Nacional de Estadística e Informática; INEI). An administrative raster file containing information on the area of each district was obtained from the GADM Global Administrative Database (Hijmans et al., 2012).

2.3 Covariates

2.3.1 Population Density in 1993

Population density obtained from the 1993 census[1] was incorporated as a measure of population in a previous time period. When data are available, previous population metrics have obvious value for predicting current population. The population density for 1993 was calculated similarly to that for 2007.

2.3.2 Geospatial Variables

[1] Peru conducted a census in 2005, but the methodology and results were considered flawed resulting in the 2007 census

A map detailing inland roads, rivers and permanent bodies of water in Peru was downloaded from the Digital Chart of the World[2] and aggregated to the third administrative level (districts) to obtain the density of roads, rivers and permanent bodies of water in each district.

Density of inland roads is an indicator of urbanization and accessibility for each district. Proximity to transportation corridors may be especially important in the Amazon where large tracts of forest make transportation difficult.

Information on inland water was included based on the propensity that urban areas have for developing around bodies of water given that they are a valuable natural resource and method of transportation. Density of rivers may be a better predictor in countries outside the Amazon basin as the Amazon River and its tributaries may create a high density of water in fairly uninhabited provinces. Data for density of permanent bodies of water was categorized as a variable separate from density of rivers.

2.3.3 Socioeconomic Variables

While land use and topography may be important predictors of population density, they are unlikely to provide useful information for areas that have already been urbanized. A GDP index derived from the 2007 and 1993 censuses was downloaded from the INEI as an economic indicator for the analysis. Although the available GDP index is on a coarse spatial scale

[2] All data that originates from the Digital Chart of the World was made public in 2006, but has not been updated since 1992. Any significant change in urbanization (specifically density of inland roads) since 1992 is not reflected in the data.

(provincial), it may provide significant information on interannual variability not present in topographical data.

2.3.4 Satellite Derived Variables

Remote sensing provides a method for detailing landscape characteristics for each district, which may in turn be linked to the population density. Remote Sensing data collected by the MODIS Terra sensor was downloaded from the NASA Reverb portal. The two chosen characteristics were the Normalized Difference Vegetation Index (NDVI) and the daytime land surface temperature (LST).

NDVI measures the difference between reflectance in the near infrared and the visible spectrum. The chlorophyll in healthy vegetation strongly absorbs visible radiation while the plant cell structure reflects it. NDVI may therefore be used both as a measure of vegetative distribution and as an indicator of vegetative health. The difference in vegetative distribution may also provide information on patterns of topographical features in the landscape, as vegetative differences are often indicative of topography. NDVI was available as a 1 km resolution gridded product, but was aggregated to district averages using the administrative shapefile to match the resolution of the other covariates. Data were available as monthly composites. For this study a consistent month during the dry season was chosen (July).

Daytime LST may act to differentiate between the diverse land cover of Peru, which includes open water, bare soil, forested areas, rock and urban areas. The diurnal thermal signal of each category of land cover may provide insight into the potential habitability of that area. Daytime

LST can also give an indication of the heat island effect of cities for some of the smaller districts in which impervious cover is an appreciable percent of total surface area. LST was available at 1 km resolution in 8-day composites. For this study the same composite was chosen from each year (mid July to match the NDVI). LST was aggregated to the district level using the GADM Global Administrative Areas shapefile.

2.4 Data Consistency

Not all of the data from the Digital Chart of the World matched the INEI districting, although discrepancies between datasets were minor. After standardizing the data, out of 417 districts present in each year (according to the most recent INEI report) 412 districts mapped to those in the Digital Chart of the World. Districts that were omitted from the study include Jesus Nazareno, Llochegua, Huepetuhe, Majes and Kimbiri. The missing districts were due to redistricting between 1992, the creation of the Digital Chart of the World, and the 1993 census.

3.0 Model Structures

Understanding and selecting the appropriate model structure is perhaps the most important decision in the process of population modeling. The fundamental act of choosing a model structure creates a lens through which all subsequent data will be interpreted and will significantly affect the understanding of covariate influence. The most appropriate model structure often depends on the data available. In this analysis, five regression and tree-based models were chosen to explore how predictive accuracy and variable importance changes in the presence or absence of population information. The regression-based model structures include a generalized linear model (GLM), a generalized additive model (GAM), and a multivariate

248 adaptive regression spline (MARS) structure. The tree-based models include Random Forest
249 (RF) and Bayesian additive regression tree (BART). A no model alternative was also included in
250 the suite of models for reference.
251
252 The models were run twice: once with population density from 1993 included as a covariate, and
253 once with it excluded, leaving only socioeconomic and environmental covariates. These two
254 analyses are hereafter referred to as being with or without population data, although neither uses
255 current period population information to estimate population density.
256
257 *3.1 Generalized Linear Models (GLM)*
258 A GLM is a linear function of the form $E(Y) = g^{-1}(X\beta) + \varepsilon$, where Y is the vectorized form of the
259 response variable, X is the covariate matrix, β is a vector of coefficients, g is the link function
260 and ε is a vector of the normally distributed errors (Cameron and Trivadi, 2005). In this case β
261 may be interpreted as the relative influence of each variable. While the model may include a link
262 function relating the covariates to the response variable, for the purposes of this model the
263 response variable was assumed to follow a Gaussian distribution so the identity link function was
264 used.
265
266 *3.2 Generalized Additive Models (GAM)*
267 A GAM is an extension of the GLM, in which the assumption of linear relationships between
268 covariates and response variables is relaxed by replacing the link function with a nonparametric
269 smoothing function, $f(X)$, such that the form of the function becomes $Y = f_1(X_1) + f_2(X_2) + \ldots +$
270 $f_n(X_n) + \varepsilon$ (Hastie et al., 2009). In this case a cubic spline was chosen for the nonparametric

smoother with restricted degrees of freedom. In this way the GAM allows for non-linear relationships between the covariates and response variables.

3.3 Random Forest (RF)

Tree-based methods are often most useful for models that are highly non-linear. The most basic tree-based structure is the Classification and Regression Tree (CART), which recursively partitions the data into i subspaces and applies a very simple model to each subspace. If the loss measure used is the sum of squares, the model takes the form $C_i = \text{mean}(Y_i|x_i \in R_i)$ where C_i is the variable to be predicted in subspace R_i, Y_i is the set of response variables on which the model is trained in that subspace and x_i is the matrix of the associated covariates. One downside of CART is that the hierarchical nature of the model means relatively small changes in the data set can result in drastically different partitions within the data space, which makes drawing insight from the model structure difficult. One approach to reduce the variability inherent in predictions from CART models is to use model averaging based on bootstrapping, a method known as bagging (Hastie et al., 2009). The RF model structure is similar to a bagged CART method, except that a random subset of variables less than the total number of variables are chosen to use at the splitting point for each tree. This produces uncorrelated trees (although not perfectly uncorrelated) such that the aggregate result is a reduction in the variance (Breiman, 2001).

3.4 Multivariate Adaptive Regression Splines (MARS)

MARS is an extension of the generalized linear class of models that allows for nonlinearity in the relationship between covariates and response variable by way of multiple basis functions that take the form $(x-t)_+$ or $(t-x)_+$ where t is a "knot point" determined in the model training process

294 and x is the covariate. The model first enumerates basis functions to fit the data and then prunes
295 back these functions, as would a tree-based model (Friedman, 1991). This gives the model the
296 form $Y = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$, where $h_m(X)$ is a basis function, or product of basis functions, and
297 β_m are coefficients estimated by minimizing the sum-of-squares.
298
299 *3.5 Bayesian Additive Regression Tree (BART)*
300 The BART model is a "sum-of-trees" method taking the form of $Y = f(X) + e$, where $f(X)$ is the
301 sum of many trees. The model places a prior probability on the nodes of each tree such that the
302 tree is constrained to be a "weak learner" (i.e. biasing the tree towards a shallower, simpler
303 structure). This constraint ensures that each tree contributes only minimally to the overall fit. The
304 model is designed to produce a flexible inference of an unknown $f(X)$.
305
306 *3.6 Mean Model*
307 Each of the previously described models was compared against the no-model mean alternative.
308 The no-model mean estimate was simply calculated as the mean of available response data in the
309 holdout dataset.
310
311 *3.7 Model Evaluation*
312 The predictive accuracy of each model, both with and without population information, was
313 evaluated using 100 repetitions of a random10-fold cross validation (CV) holdout analysis. Only
314 one year of census data was available for the response variable dataset, meaning that every
315 prediction made in the CV analysis was an out-of-sample prediction. This strengthens the results

316 of the analysis for predictions made in 2007, but leaves the generalization of these results to
317 other time periods untested.
318
319 The suite of models is assessed using mean absolute error (MAE) as a measure of general model
320 accuracy as well as root mean square error (RMSE). The difference between MAE and RMSE is
321 used to assess the skill of each model in providing population density estimates for the highest
322 density districts. In evaluating the results of each analysis, reported levels of statistical
323 significance are measured to a significance level of 0.05 following a Bonferroni correction for
324 multiple pair-wise comparisons and are based on MAE (See Tables 1 and 2).
325
326 The diversity of model structures included in the analysis required the use of multiple measures
327 of variable influence in the analysis of the results. The relative importance of each variable in a
328 GLM was measured using the β coefficient from the final fitted model. In this case the β
329 coefficient indicates the linear relationship between covariate and response variable (see Sect.
330 3.1). Variable influence in the MARS model was based on the contribution of a variable towards
331 reducing the model's generalized cross-validation (GCV) score. GCV is an approximation of the
332 leave-one-out cross-validation using a squared error loss measure (Hastie et al., 2009). The
333 measure of variable importance used for a GAM was the increase in MSE that results from
334 removing a specific variable. Variable importance in the RF model was evaluated using two
335 separate indices. The first is based on perturbing each variable and recording the effect on the
336 out-of-bag accuracy as measured by MSE, while the second measures the decrease in node
337 impurities- measured by the residual sum of squares- that results from splitting on a variable.
338 Variable importance in the BART model was evaluated by the number of times a variable was

used as a splitting decision in a tree, averaged over all trees. Due to the discrepancy between measures of variable influence, direct comparisons between models cannot be made. Instead, the shift in relative variable influence between analyses is explored within each model to understand how each model is affected by the presence of population density information in the covariates.

4.0 Results and Discussion

The differences in model accuracy as evaluated by RMSE as opposed to MAE are minimal in terms of ordinal rank but entail consistently larger mean error estimates with increased standard errors. The systematic difference in model accuracy as measured by MAE and RMSE (results not shown) implies that the skill of each model varied significantly as a function of the districts chosen for the holdout sample and in fact that a minority of the holdout samples were driving the model estimate errors, a result that most likely stems from the out-of-sample nature of predictions imposed by the limited dataset. An expanded dataset - either a greater number of districts or greater number of years – may help to reduce the standard error of the model estimates.

Despite the fact that the regression based models (GLM, GAM and MARS) provided the most skilled predictions of population density when 1993 population density was included in the covariates, these models provided among the worst predictions when no population information was included (see Tables 1 and 2). In fact, when population density data was not included, none of the regression based models produced predictions that performed better than the no-model alternative (Table 2). This result indicates that when population information was not available

361 regression based models were unable to capture the relationship between indicator variables and
362 current population density.
363
364 In contrast to the regression models, the RF model – a non-parametric tree based model -
365 provided among the most skilled estimates when 1993 population density was not included in the
366 covariates, but among the least skilled estimates when it was (see Tables 1 and 2). Notably, the
367 RF model was the only model to significantly outperform the no-model alternative when
368 population information was not included. The shift in relative model performance indicates that
369 the relationship between previous population density and current population density at a district
370 scale can be modeled effectively using regression methods, but the relationships between
371 ancillary variables and population density require a non-parametric model structure due to either
372 nonlinearity or a large degree of variability.
373
374 The covariate influence of all models was explored to understand the differences in variable
375 importance between the two analyses. Although the most direct measure of variable importance
376 is model dependent, which precludes direct comparisons between measures of variable
377 importance, relative comparisons between analyses are possible and instructive. When
378 population density from 1993 was included in the covariates, GLM, GAM and MARS – the three
379 models that provided the best population density estimates for the analysis– all indicated that
380 previous population density was the most significant variable as assessed by their respective
381 measures of variable importance (Table 3). This relative variable importance is unsurprising in-
382 and-of itself, but is an important point of comparison for evaluating the models that do not
383 include 1993 population density information in the covariates.

When 1993 population information was not included in the analysis, nearly all of the models incorporated a greater number of covariates, many of which the models had previously excluded completely (see Tables 3 and 4). Random Forest – the model that provided the best population density estimates for the analysis - indicated that the majority of remaining covariates had comparable variable importance (Table 4). The RF model therefore compensated for a lack of previous population density information more effectively than regression-based models by incorporating information from nearly all of the available covariates.

Random Forest population density estimates and model errors are explored spatially and in their relation to actual district population density to better understand the performance of the model. Figure 2 demonstrates minimal spatial dependencies in the model errors with mixed performance in the mid-latitudes and a consistent overestimation of population density to the south in the region of Arequipa, particularly those districts surrounding the city of Arequipa. Figure 3 shows that RF systematically underestimated the population density of the highest-density districts and tended to overestimate population density of the mid- to low-density level districts. The overestimation bias for low-density districts is not surprising given the relatively small margin available for underestimation in such districts. The inability of the RF model to produce accurate population density estimates for the most population dense districts implies that the resolution of the analysis – which in this case is the district level – may have been insufficient to capture the upper extreme of population-density due to heterogeneity of the response variable within each district. Dense urban areas may account for the majority of a district's population but a relatively minimal proportion of its land area, on which many covariates were based.

5.0 Conclusions

The presence or absence of population information drastically changes the problem of population modeling from one that may be accurately modeled with only one or two variables using regression methods into one that benefits from multiple covariates in a non-parametric model structure. The intuitive implication of these results is that the appropriate model structure is dependent on the quantity and quality of previous population information. Somewhat less intuitively, the results demonstrate that population density estimates can be made for regions lacking previous population information altogether by training a non-parametric tree-based model on data from culturally and demographically similar areas. Such estimates are vital for decision makers operating in regions limited by incomplete or unreliable census data.

The effort to create frequently updated, spatially distributed population estimates in data-limited regions is driving population modeling towards including an increasing number of variables. Current operational inter-census statistical models often either use demographic data or a single variable land-use classification scheme to estimate population densities. These methods may provide good results for projecting stable population growth, but will perform less well in areas of new development or those lacking reliable census data. For regions in which data limitations preclude the use of reliable demographic information, it is important that model structures effectively incorporate all relevant ancillary data. This paper demonstrates that for the five chosen regions in Peru there exists a stark difference between the appropriate model structure dependent upon the presence or absence of reliable census data. The predictive accuracy of tree-

429 based non-parametric models in population modeling is an area that has been largely unexplored
430 but which may yet prove tremendously useful for estimating population density.

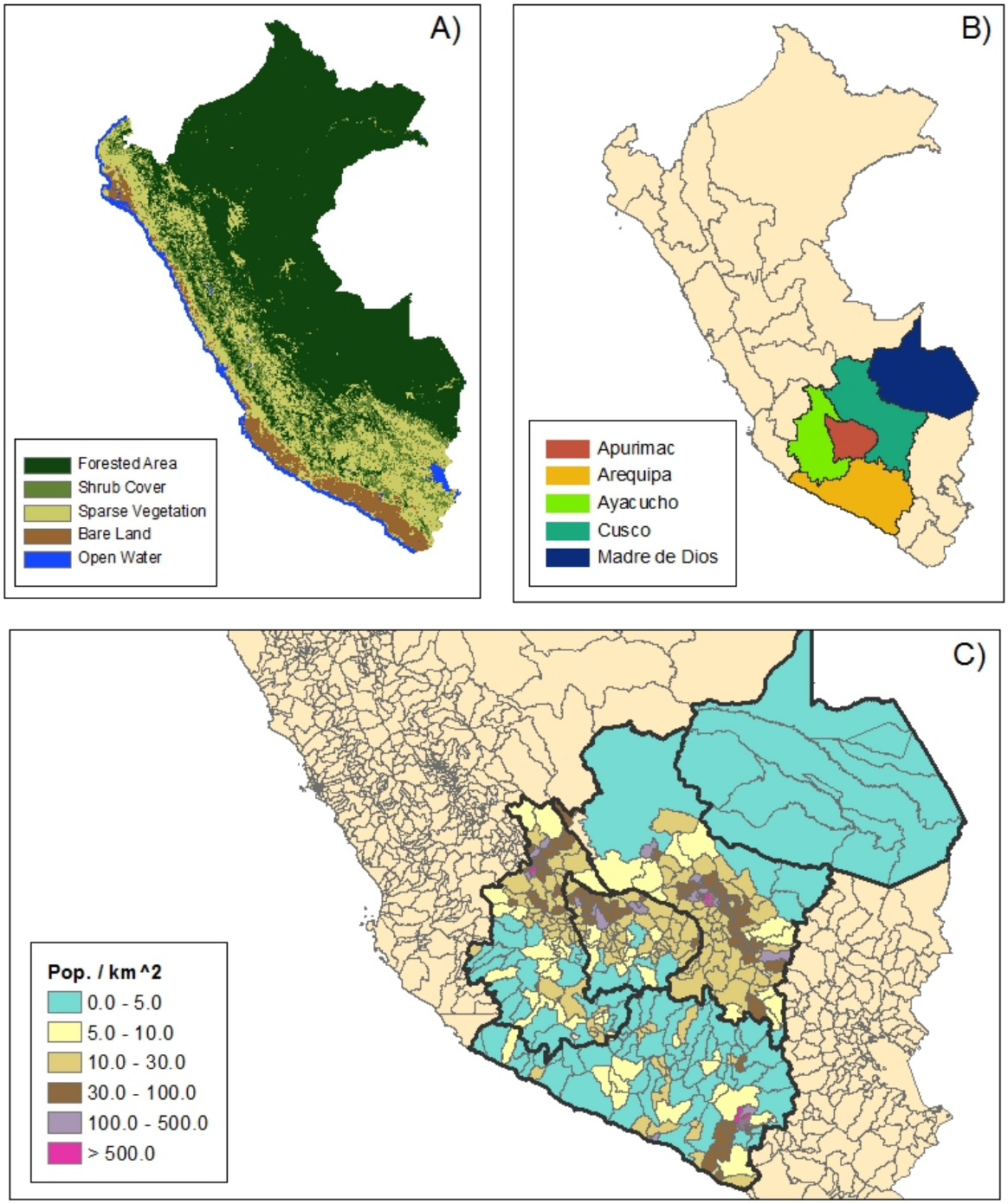

Figure 1: Political boundaries and topographical features of Peru. A) Classified land cover. B) Regions included in the analysis (Apurimac, Arequipa, Ayacucho, Cusco and Madre de Dios). C) Population density at the district level derived from the 2007 census.

Table 1: Model errors and p-values for models that included population data in the covariates. Shading indicates p-values corresponding to the t-test between the MAE distributions of each row-column pair[a]. Orange (p-value > 0.10), Yellow (0.05 < p-value < 0.10) and Green (p-value < 0.05)

	Mean Absolute Error	Standard Error	GLM	GAM	MARS	RF	BART	Mean
GLM	0.063	0.032	1.00E+00					
GAM	0.057	0.054	1.00E+00	1.00E+00				
MARS	0.056	0.039	1.00E+00	1.00E+00	1.00E+00			
RF	0.104	0.086	2.72E-04	1.07E-04	1.47E-05	1.00E+00		
BART	0.079	0.051	1.29E-01	4.72E-02	4.95E-03	2.14E-01	1.00E+00	
Mean	0.314	0.251	2.08E-15	7.42E-16	4.83E-16	2.15E-11	6.59E-14	1.00E+00

[a] For example, the average MAE of the BART population estimates for this analysis is 0.079 with a corresponding standard error of 0.051. The MAE distribution for the BART model is statistically significantly distinct from GAM, MARS and the Mean model but is not statistically distinct from the GLM or RF MAE distributions.

Table 2: Model errors and p-values for models that did not include population data in the covariates. Shading indicates p-values corresponding to the t-test between the MAE distributions of each row-column pair. Orange (p-value > 0.10), Yellow (0.05 < p-value < 0.10) and Green (p-value < 0.05)

	Mean Absolute Error	Standard Error	GLM	GAM	MARS	RF	BART	Mean
GLM	0.372	0.118	1.00E+00					
GAM	0.381	0.117	1.00E+00	1.00E+00				
MARS	0.339	0.139	1.00E+00	3.30E-01	1.00E+00			
RF	0.207	0.116	3.83E-18	7.65E-20	1.07E-10	1.00E+00		
BART	0.289	0.158	5.19E-04	7.70E-05	2.62E-01	6.99E-04	1.00E+00	
Mean	0.314	0.251	5.34E-01	2.42E-01	1.00E+00	2.62E-03	1.00E+00	1.00E+00

Table 3: Measures of variable importance when population data is included in the covariates. Missing numbers indicate variables that were discarded by the model during variable selection. Shading indicates the models producing the most accurate estimates.

	Previous Popdensity	Roads	River Water	X coordinate	Y coordinate	NDVI	LST Day	GDP	Perm Water
GLM Beta Values	0.979	-	-0.0199	-	-	-	-	-	-
GAM Percent reduction in MSE	505.22	-0.92	-0.87	-0.07	1.60	1.78	-	-	-
MARS GCV	100	5	-	-	-	-	-	-	-
RF Percent reduction in MSE	22.19	1	2.26	1.35	3.9	3.33	3.74	2.15	1.21
RF Inc. Node Purity	204.94	81.85	22.21	16.25	27.27	19.65	18.34	3.11	0
BART Mean number of splits	68.28	19.89	13.84	11.95	10.76	11.71	31.21	17.88	27.29

Table 4: Measures of variable importance when population data is not included in the covariates. Missing numbers indicate variables that were discarded by the model during variable selection. Shading indicates the models producing the most accurate estimates.

	Previous PopDensity	Roads	River Water	X coordinate	Y coordinate	NDVI	LST Day	GDP	Perm Water
GLM Beta Values	NA	0.154	-0.184	0.117	-0.102	-	-	-	-
GAM Percent reduction in MSE	NA	-4.37	-1.30	-	-2.33	-2.59	-	-	-
MARS GCV	NA	100	45.9	15.8	26	22.2	16.8	-	-
RF Percent reduction in MSE	NA	3.19	-0.91	2.45	4.7	4.88	5.61	4.33	-0.63
RF Inc. Node Purity	NA	116.79	45.36	35.23	47.67	32.61	41.14	7.78	0.18
BART Mean number of splits	NA	55.03	28.43	18.09	33.01	30.33	39.62	37.19	26.61

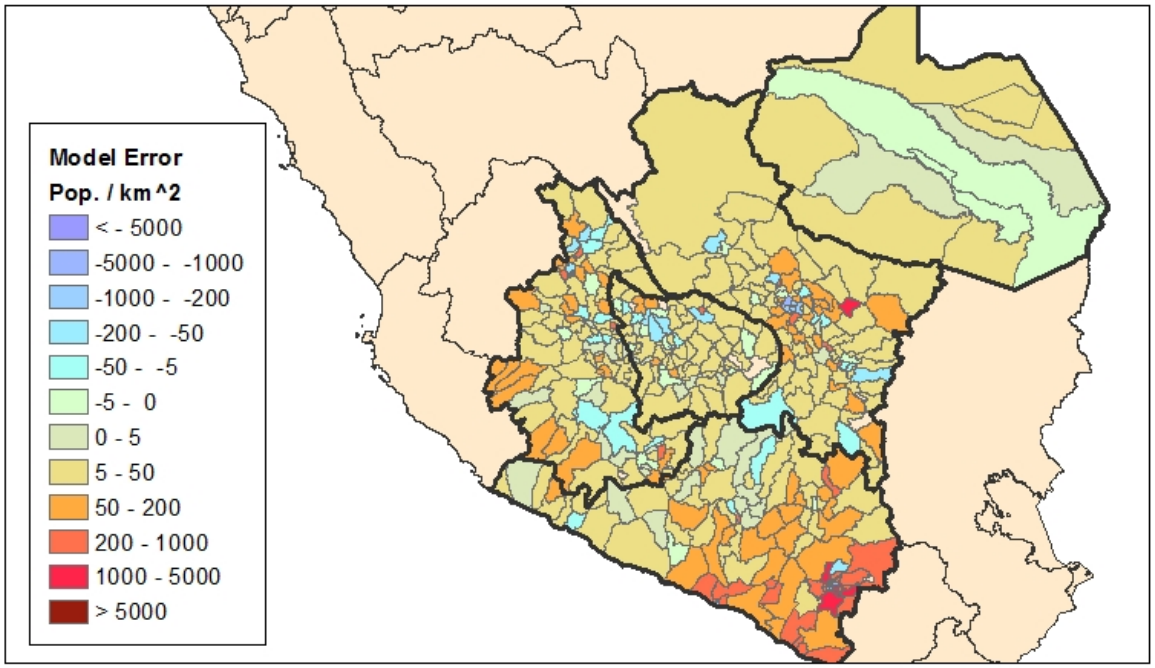

Figure 2: Random Forest model error by district

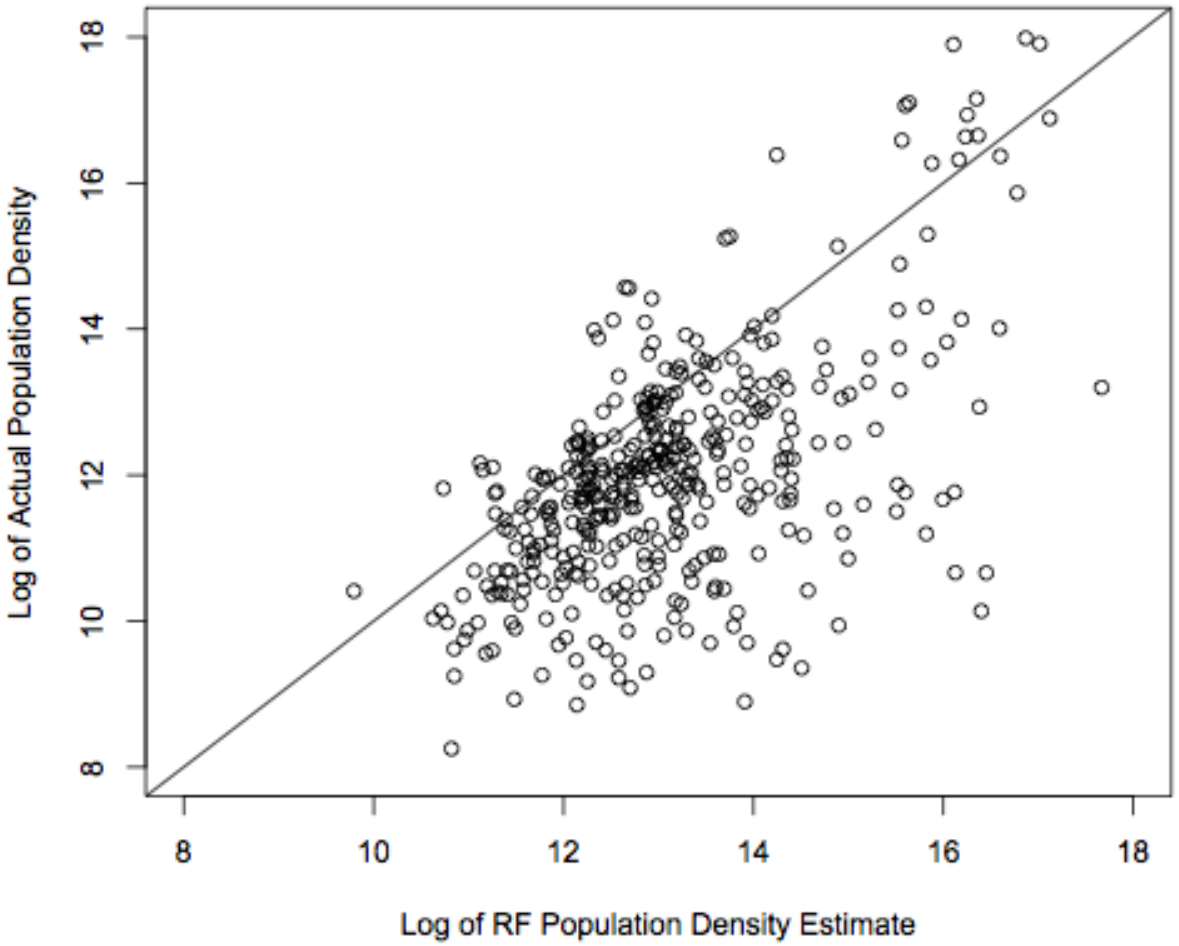

Figure 3: Actual population density vs. RF estimated density with a 1:1 line plotted for reference

Works Cited

Balk, D., Brickman, M., Anderson, B., Pozzi,. F., Yetman, G. (2005). Annex, Estimates of future gloval population distribution to 2015. Food and Agriculture Organization of the United Nations (FAO).

Boudot, Y (1993). Application of remote sensing to urban population estimation: a case study of Marrakech, Morocco. EARSeL Advances in Remote Sensing, Vol 3, No 3.

Breiman, L. (2001). Random forests, Machine Learning, Vol 45(1), No 5-32.

Cameron, A.C. and Trivedi, P.K. (2005). Microeconometrics: Methods and Applications. Cambridge: Cambridge University Press.

de Castro, M.C., Sawyer, D.O., Singer, B.H. (2007). Spatial patterns of malaria in the Amazon: implications for surveillance and targeted interventions, Health & Place, Vol 13, pp 368-380.

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., and Worley, B.A. (2000). LandScan: A Clogal Population Database for Estimating Populations at Risk, Photogrammetric Engineering & Remote Sensing. Vol 66, No 7.

Eicher, C. and Brewer, C. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. Cartography and Geographic Information Science, 28, 125–138.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. Annals of Statistics 19 (1): 1–67. doi:10.1214/aos/1176347963.

Hastie, T., Tibshirani, R., Friedman, J. (2009). Additive Models, Trees and Related Methods. Chapter 9 of The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag.

Harvey, J. T. (2002). Population Estimation Models Based on Individual TM Pixels, Photogrammetric Engineering and Remote Sensing, 68(11): 1181 – 1192.

Hijmans, R., Garcia, N., Rala, A., Maunahan, A., Wieczork, J. and Kapoor, J. (2012). GADM Global Administrative Areas Version 2, www.gadm.org.

Langford, M. Maguire, D. J. and Unwin, D. J. (1991). "The Aerial Interpolation Problem: Estimating Population Using Remote Sensing in a GIS Freamework" in Handling Geographical Information: Methodology and Potential Applications, Masser, I. and Blakemore, M. (Eds), New York, NY: Wiley, 55-77.

Liu, X., Clarke, K.C. (2002). Estimation of Residential Population Using High Resolution Satellite Imagery, Proceedings of the 3rd Symposium in Remote Sensing of Urban Areas, Istanbul, Turkey. June 11-13.

Lo, C.P. (2003). Zone Based Estimation of Population and Housing Units from Satellite-Generated Land Use/Land Cover Maps. In V. Mesev (Ed.), Remotely Sensed Cities (pp. 157-180). London, UK/New York, NY: Yaylor & Francis.

O'Neill, B., Balk, D., Brickman, M., Ezra, M. (2001). A Guide to Global Population Projections. Demographic Research, 4(8) p203-288. doi: 10.4054/DemRes.2001.4.8.

Puissant, A. (2010). Estimating population using remote sensing imagery. Panel contribution to the Population-Environment Research Network Cyberseminar, "What are the remote sensing data needs of the population-environment research community?"

Sanderson, W. C. (1998). Knowledge Can Improve Forecasts: A Review of Selected Socioeconomic Population Projection Models. Population and Development Review, 24 p88-117.

Wu, S., Qiu, X., Wange, L. (2005). Population Estimation Methods in GIS and Remote Sensing: A Review. GIScience and Remote Sensing, 42 p58-74.