COMPILING GRANULAR POPULATION DATA USING GEOSPATIAL INFORMATION

Thomas Mitterling, Katharina Fenz, Arturo Martinez Jr., Joseph Bulan, Mildred Addawe, Ron Lester Durante, and Marymell Martillan

NO. 643

December 2021

ADB ECONOMICS
WORKING PAPER SERIES

ASIAN DEVELOPMENT BANK

ADB Economics Working Paper Series

Compiling Granular Population Data Using Geospatial Information

Thomas Mitterling, Katharina Fenz, Arturo Martinez Jr., Joseph Bulan, Mildred Addawe, Ron Lester Durante, and Marymell Martillan

No. 643 | December 2021

Thomas Mitterling and Katharina Fenz are both senior data scientists at the World Data Lab. Arturo Martinez Jr. is a statistician; Joseph Bulan is an associate statistics analyst; and Mildred Addawe, Ron Lester Durante, and Marymell Martillan are consultants at the Economic Research and Regional Cooperation Department, Asian Development Bank.

ASIAN DEVELOPMENT BANK

6 ADB Avenue, Mandaluyong City, 1550 Metro Manila, Philippines
Tel +63 2 8632 4444; Fax +63 2 8636 2444
www.adb.org

ISSN 2313-6537 (print), 2313-6545 (electronic)
Publication Stock No. WPS210519-2
DOI: http://dx.doi.org/10.22617/WPS210519-2

The views expressed in this publication are those of the authors and do not necessarily reflect the views and policies of the Asian Development Bank (ADB) or its Board of Governors or the governments they represent.

ADB does not guarantee the accuracy of the data included in this publication and accepts no responsibility for any consequence of their use. The mention of specific companies or products of manufacturers does not imply that they are endorsed or recommended by ADB in preference to others of a similar nature that are not mentioned.

By making any designation of or reference to a particular territory or geographic area, or by using the term "country" in this document, ADB does not intend to make any judgments as to the legal or other status of any territory or area.

Corrigenda to ADB publications may be found at http://www.adb.org/publications/corrigenda.

Notes:
In this publication, "$" refers to United States dollars.
ADB recognizes "USA" as the United States.

The ADB Economics Working Paper Series presents data, information, and/or findings from ongoing research and studies to encourage exchange of ideas and to elicit comment and feedback about development issues in Asia and the Pacific. Since papers in this series are intended for quick and easy dissemination, the content may or may not be fully edited and may later be modified for final publication.

CONTENTS

TABLES AND FIGURES

TABLES

FIGURES

ABSTRACT

Granular spatial information on the distributions of human population is relevant to a variety of fields like health, economics, and other areas of public sector planning. This paper applies ensemble methods and aims at assessing their applicability to analyzing and forecasting population density on a grid level. In a first step, we use a Random Forest approach to estimate population density in the Philippines and Thailand on a 100 meter by 100-meter level. Second, we use different specifications of Random Forest and Bayesian model averaging techniques to create forecasts of the grid-level population density in three Thailand provinces and evaluate their predictive power.

Keywords: population mapping, big data, random forest estimation, Philippines, Thailand

JEL codes: C19, D30, O15

I. INTRODUCTION

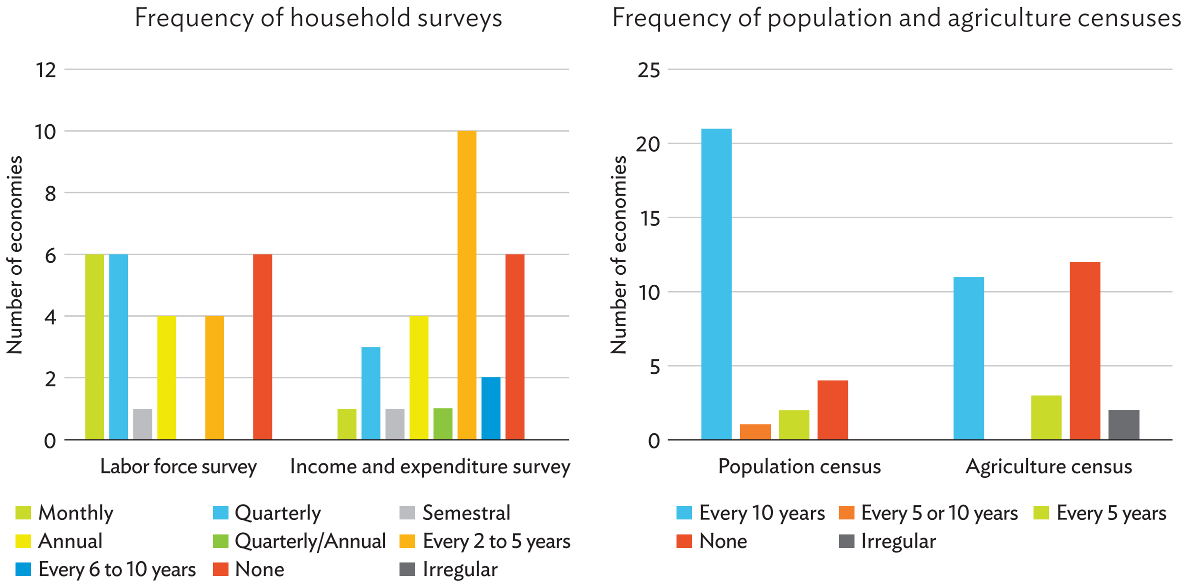
To appreciate the importance of having accurate and timely data on distributions of human population, it is important to understand how they are used. Population numbers inform a wide range of areas of decision-making and help governments assess where resources and investments need to be allocated (Balk et al. 2006; Salvatore et al. 2005; Tatem et al. 2012). With population size continuing to be very dynamic over the next decades, accurate data, and forecasts of population distributions are becoming even more important (UN DESA 2018). With spatially disaggregated population datasets, governments can identify areas that need new roads, schools, hospitals, and other infrastructure and service facilities. The importance of population data is further highlighted in the 2030 Sustainable Development Agenda where population counts are integral for monitoring a wide array of indicators. In addition, population counts are required in the context of assessing the situation of people by income, age, sex, ethnicity, migratory and disability status, and geographic area. Accurate and updated population numbers are also fundamental in planning intervention programs in response to disasters and other economic shocks. This was recently demonstrated when the coronavirus disease pandemic struck as countries and economies heavily relied on population data not only to identify potential hotspots of widespread infection, but to determine how much relief efforts are warranted for specific population groups and geographic areas.

Given the importance of having accurate, timely, and granular demographic data, population numbers constitute one of the basic types of data collected by national statistics offices or other government agencies mandated to collect data that are useful inputs for policy planning. There are different channels through which population data are compiled. In many developed countries, population data are extracted from national registers of births and deaths. The accuracy of information recorded in national registers, synchronization from local to national-level registry systems, as well as frequency of updating of data recorded therein determine the quality of population numbers that can be extracted from national registers. National population registers, when managed properly, have the advantage of producing near real-time population data.

In addition to national registers, population and housing censuses are also considered one of the commonly used data collection vehicles to compile population numbers. In fact, the World Population and Housing Census Programme is one of the longest-standing global statistical programs, which recognizes the importance of household and population census as an important source for supplying data on distributions of human population. National statistical offices or similar government agencies differ in modes of conducting population and housing census and each of these modes has strengths and weaknesses in maximizing response rates in a cost-effective manner, while maintaining high-quality data. Traditional approaches rely on face-to-face interviews. With well-trained interviewers, face-to-face interviews ensure common understanding of questions, thus facilitating enhanced comparability of data throughout the country. Obviously, this mode comes at a greater cost of conducting population and housing censuses. On the other hand, other countries conduct their census by sending questionnaires through postal mails and giving people the option of responding by mail or Internet. These approaches are presumably more cost-effective but could be more prone to systematic nonresponse patterns. Telephone interviews are also practiced in other countries.

Despite the availability of ways to minimize the cost of conducting population and household censuses, the fact remains that financially constrained countries are unable to conduct a census frequently. On average, population and household censuses are conducted every 10 years. In Asia and the Pacific, less than 10 countries conduct a census more frequently than every 10 years as the average per capita cost of conducting one is around $2.04 (SDSN 2015). This corroborates the findings of a survey conducted by the Statistics and Data Innovation Unit (SDIU) of the Asian Development Bank (ADB) which show that a nonnegligible portion of its regional economies conduct major surveys and censuses less frequently than ideal (Figure 1) despite improvements in overall statistical capacity over time (Figure 2).

Figure 1: Frequency of Surveys and Censuses in Developing Asia and Pacific Economies



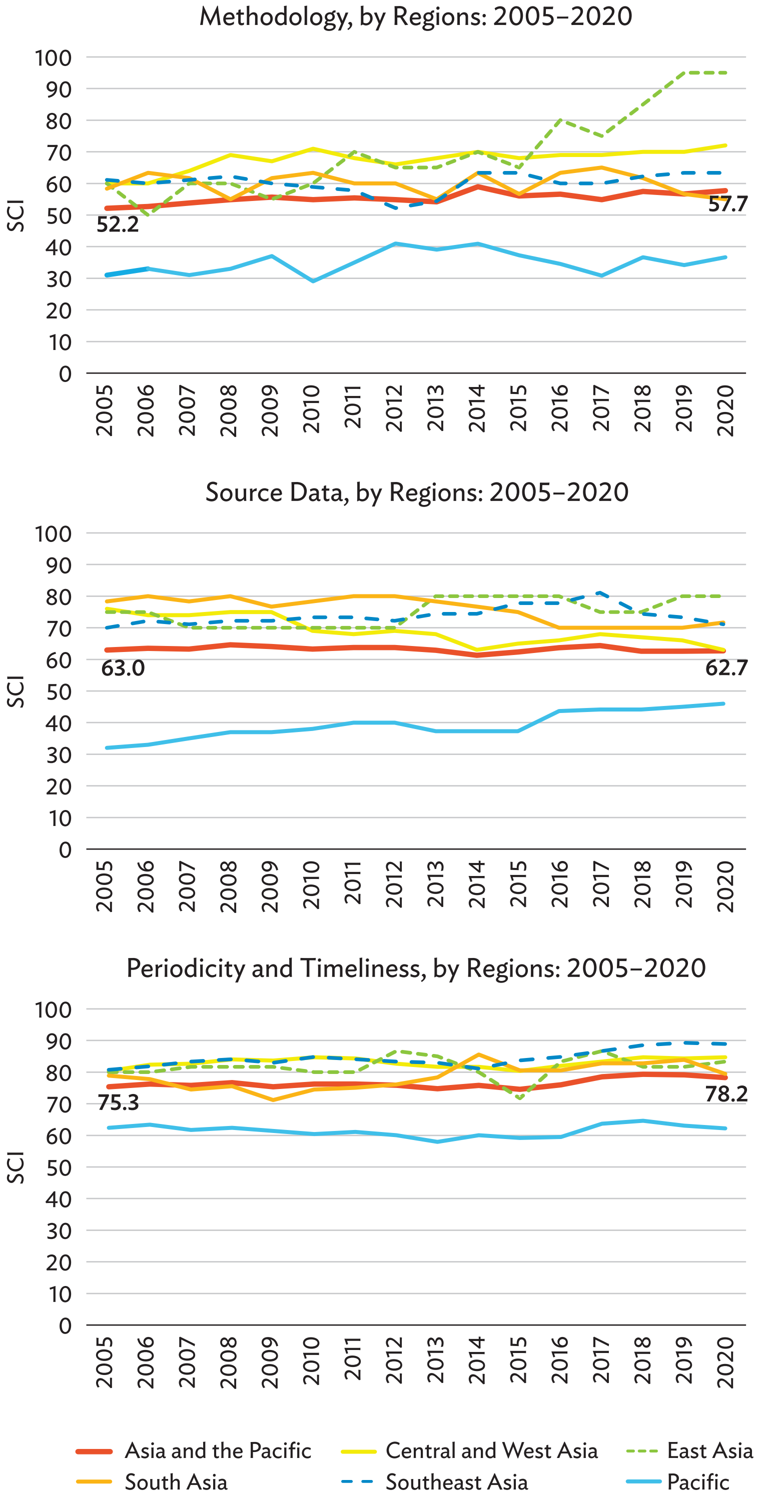
Note: The analysis uses results from 28 regional member economies of the Asian Development Bank.

Source: Adapted from ADB. 2021. *Key Indicators for Asia and the Pacific 2021*. Manila. (Figure 4.1).

The coronavirus disease pandemic further intensified the challenge of conducting timely census. There are 120 countries or economies worldwide that have either postponed, delayed or canceled the conduct of their respective population censuses (UNCTAD 2021). Some of these countries changed their data collection methodologies to push through with their census activities. In the Asia and Pacific region, 6 out of the 10 member economies surveyed by SDIU rescheduled their census field activities to either late in 2020 or into 2021. National statistics offices also intensified initiatives to capacitate their staff and enumerators on the use of new methods to facilitate timely census data compilation.

Representativeness of population data is another area where potential enhancements could be made. Accuracy and quality of population estimates for remote or conflict-stricken areas might be subject to criticisms as it is challenging to collect data from these areas, especially when face-to-face interviews are the chosen mode of data collection. Granularity could also be improved as the extent to which population data can be spatially disaggregated is constrained by the range of geographic information included in the data collection instrument. However, as research and decision-making become progressively complex, the demand for more timely, representative, and granular population data increases. To meet such data requirements, alternative methods of compiling population data are being explored. These methods are diverse, but many of them exploit data sources that are not traditionally used for population estimation.

Figure 2: Statistical Capacity Indicator in the Asia and Pacific Region



Note: The analysis uses data from ADB regional member economies for which estimates of the Statistical Capacity Indicator (SCI) are available.

Source: Adapted from ADB. 2021. *Key Indicators for Asia and the Pacific 2021.* Manila. (Figure 4.2).

One major resource in the creation of very granular population maps is data based on satellite imagery. The use of remotely sensed data is already well established in the research on population density and has proved especially important in countries with a limited availability of recent and accurate census data. However, even in countries with rich and reliable information from censuses and surveys, satellite imagery plays a central role in bringing estimations to an even more granular geographical level (Tatem et al. 2007). That is why a considerable number of studies have already applied remotely sensed data as their main source of input data, or have used it in combination with census information (Anderson and Anderson 1973; Bhaduri et al. 2007; Chen 2002; Linard, Gilbert, and Tatem 2011; Linard et al. 2012; Sutton et al. 2001). Besides its granularity, another advantage of satellite imagery is that it can even be updated when interventions like face-to-face interviews are not feasible. Hence, in exceptional situations like during a pandemic, this source becomes even more relevant.

Often, most initiatives to compile grid-level population density have relied on the application of basic dasymetric population mapping (Balk and Yetman 2004; Balk et al. 2006; Tobler et al. 1995). Dasymetric population mapping entails dividing administrative areas into smaller spatial units, onto which population size is averaged to calculate population density. The Gridded Population of the World database and the Global Rural Urban Mapping Project are examples of global datasets that have been produced with this approach. To obtain their estimates, Gridded Population of the World uses a simple redistribution across census units, while Global Rural Urban Mapping Project specifically considers rural and urban areas in the spatial distribution of population inside census units. Another dataset with global coverage is WorldPop's gridded population count database. WorldPop applies an even more elaborate approach, which is based on a Random Forest estimation technique (Stevens et al. 2015). This dasymetric redistribution approach allows for the inclusion of a wide range of input data, including remotely sensed and geospatial data from different scales. Additionally, Random Forests benefit from non-parametric spatial predictions, which are then anchored across space using contemporary administrative boundary-linked geographic information system census data. This method also represents the basis for the estimations described in this paper. Datasets mapping grid-level population counts for larger areas have also been produced by LandScan as well as by the United Nations Environment Programme (Bhaduri et al. 2007; Deichmann 1996; Dobson et al. 2000). Several studies like those by Azar et al. (2010); Azar et al. (2013); and Lung et al. (2013) have also used high-resolution satellite imagery to obtain granular estimates of population density in select countries too.

More recently, Stevens et al. (2015) have proposed ensemble methods in population mapping, which capitalize on remotely sensed and geospatial data. In particular, they have developed a new semi-automated dasymetric modeling approach which incorporated detailed census and ancillary data in a flexible Random Forest estimation technique. Using Cambodia, Kenya, and Viet Nam as case studies, they were able to show that their proposed method increases over the accuracy and flexibility of other common gridded population data production methodologies.

This study applies the new method proposed by Stevens et al. (2015) to compile granular population data for two other Asian countries: the Philippines and Thailand. More specifically, we combine government-published data with publicly available data on parameters like land cover classes, elevation, slope, and nighttime lights on different levels of granularity and apply a Random Forest model to obtain grid-level estimates of population density on the 100 meters by 100 meters grid level.

In addition to compiling population data that are more granular than government-published estimates, another methodological contribution of this study is to examine the feasibility of forecasting the population distribution at the grid level using data from three provinces in Thailand as case studies. To ensure the validity of these forecasts or projections, we evaluate different model specifications and, in addition to Random Forests, also consider a Bayesian Model Averaging (BMA) approach.

The applications of such granular forecasts of population density are manifold. First, they can support the planning and administration of a new census or survey. Second, especially during a pandemic or other state of exception in which the possibility of conducting face-to-face interviews might be limited, forecasts can complement census data and facilitate updating outdated information. Third, granular forecasts of population density can show where help is needed the most in situations of crisis or catastrophe, especially when changes in people's behavior can be adequately captured by changes in spatial features on the ground. Lastly, projections like these may also be very valuable to determine the best locations for new facilities like schools, hospitals, and other public amenities.

The remainder of this paper is structured as follows: Section II gives an overview of all input data used; Section III describes the methods applied; Section IV provides details on the results; and Section V concludes the paper.

II. DATA

A. Data Sources

We use population data compiled by the Philippine Statistics Authority and National Statistical Office of Thailand. Country-specific published population data were matched to geographic information system-delineated administrative boundaries for the municipal and *barangay* (village) level in the Philippines and *tambon* (township) level in Thailand. These administrative levels provide the most granular level administrative unit available at the time of analysis. To estimate population density at the grid level, the natural logarithm of population density is used as the response variable. This decision is consistent with the recommendation of Stevens et al. (2015) who experimented with different specifications of the response variable, including the natural logarithm, the common logarithm, and the square root of population density; and found out that using the natural logarithm yields the highest prediction accuracy. On the other hand, to forecast future levels of population density for select provinces in Thailand, we use the growth of the population density as the dependent variable.[1]

Population distribution is usually highly correlated with land cover types, so we incorporated land cover information using GlobCover, a global land cover map based on Environmental Satellite's Medium Resolution Imaging Spectrometer. This global land cover map counts 22 land cover classes defined within the United Nations (UN) Land Cover Classification System and is available at a spatial resolution of approximately 300 meters. In particular, 21 of these 22 land cover classes can be found in Thailand, while 12 of them can be found in the Philippines. All those classes are included in our estimations. We use GlobCover as our source of information on different land cover classes since GlobCover provides high-quality, publicly available data and good coverage for both countries of interest, as demonstrated in previous studies (Stevens et al. 2015).

[1] In addition to providing more granular population estimates, we also examined the feasibility of a forecasting methodology to project the population size in 2020 of select provinces in Thailand. This exercise was limited to the provinces of Udon Thani, Uthai Thani, and Samut Songkhram, which were strategically identified with the assistance of the National Statistical Office of Thailand. At the time of preparing this study, the main objective was to compare the projections with the data to be collected from the 2020 Census of Population and Housing, which was scheduled in the first semester of 2020. However, due to the pandemic, the census field operations had to be postponed.

Land cover data is complemented by digital elevation data and its derived slope estimate based on HydroSHEDS data. Furthermore, we include MODIS-derived net primary production, monthly data on temperature and precipitation from WorldClim, and information on nighttime lights from the Visible Infrared Imaging Radiometer Suite. Finally, we use OpenStreetMap to obtain information on different features like villages, schools, and rivers, and Protected Planet to identify protected areas.

The table presented in the appendix gives an overview of the input data used to estimate population density on the 100 meters by 100 meters grid level. To put it into perspective, an average-sized *tambon* in Thailand is about 70 square kilometers, so there are about 7,000 grids inside each *tambon*. On the other hand, an average-sized *barangay* in the Philippines is about 7 square kilometers, so there are around 700 grids inside each *barangay*. Note that most of the input data have a constant resolution in degrees, and therefore the resolution in meters changes with the distance to the equator. The resolution shown in the table refers to the approximate resolution near the equator.

B. Data Processing

Before applying Random Forest and BMA estimations to the data, we transformed them into the same raster format. For vector data this means rasterization, while in case of raster data the projection, resolution, and origin must be changed to get consistent raster files.

We have two sources of input data vector. The first one is the data set on the protected areas, which consists of polygon and point shapefiles. The second one is OpenStreetMap data set wherein the OpenStreetMap files contain line, point, and polygon attributes. We brought all these raster data into the same projection. To convert them into raster data, we then rasterized all points, lines, and polygons. In instances when data from OpenStreetMap were not available for the target year, we used the closest year for which data are available.

III. METHODOLOGY

A. Random Forest[2]

To estimate population density on a grid level, a Random Forest approach is applied. As briefly mentioned earlier, the Random Forest is also one of the two techniques that we assessed for forecasting population density. The second approach, Bayesian model averaging (BMA), is explained in more detail in the second part of this section.

Random Forests are an ensemble method based on trees, with each tree building on a random subset of the training data and a random subset of the independent variables (Breiman 2001; Cutler, Cutler, and Stevens 2012). In particular, it is assumed that the p-dimensional vector

[2] There are studies that use principal components analysis (PCA) before fitting a ridge regression model. As PCA reduces the dimension of data, it enhances computational time which is one of the main challenges for Random Forest when working with a large set of covariates. However, introduction of PCA makes the interpretation of significant covariates and features more complicated. In this study, we immediately estimated ridge regression to facilitate a more intuitive interpretation of results. Nevertheless, future studies can also include PCA to examine robustness of results.

$X = (X_1, \dots, X_p)^T$ of explanatory variables and the dependent variable Y follow a joint distribution $P_{XY}(XY)$. In this paper, Y is the number of persons living in each 100 meters by 100 meters grid cell and X consists primarily of geospatial data like nightlights, land cover classes, temperature, and precipitation.

The aim of a Random Forest model is to find a function $f(X)$ to predict Y. This prediction function is determined by a loss function $L(Y, f(X))$ and should minimize the expected loss:

$$E_{XY}(L(Y, f(X))) \tag{1}$$

with respect to the joint distribution of X and Y.

The loss L is determined by how close $f(X)$ is to Y. The farther $f(X)$ is from the observed value, the higher the experienced loss. When Y is a continuous variable estimated with a Random Forest of regression trees, L is defined as the squared error loss $L\big(Y, f(X)\big) = (Y - f(X))^2$. In this case, minimizing the expected value of the squared error loss $E_{XY}\left(L\big(Y, f(X)\big)\right)$ leads to the regression function:

$$f(X) = E(Y|X = x) \tag{2}$$

A Random Forest of this form is created from a set of regression trees $h_1(x), \dots, h_J(x)$ as base learners. Averaging over the predictions of all individual trees leads to the Random Forest prediction:

$$f(X) = \frac{1}{J}\sum_{j=1}^{J} h_j(x) \tag{3}$$

(1) Regression Trees

Since every tree only takes into account a random subset of explanatory factors, the jth tree can be denoted as $h_j\big(X, \Theta_j\big)$, where Θ_j is a random subset of the covariates and the Θ_js are independent for $j = 1, \dots, J$. Considering a set of training data $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$, where $x_i = (x_{i,1}, \dots, x_{i,p})^T$ includes p covariates, y_i is the response, and θ_j is a particular realization of Θ_j, a fitted tree can be written as $h_j(x, \theta_j, D)$ (Breimann 2001).

However, the parameter θ_j, which adds randomness to the trees, is not modeled explicitly, but enters the trees implicitly in two different ways. First, in the splitting of each node, only a random subset of all covariates is considered. Second, each tree only takes into account a bootstrap sample of the training data. This means that for each tree a random subset of the training data is drawn with replacement (Cutler, Cutler, and Stevens 2012).

The regression trees creating a Random Forest are grown using the method of recursive binary splitting. In other words, each tree partitions the predictor space in a sequence of binary splits based on individual covariates. The first node, including the entire predictor space, is called the root node. The final nodes that are not split any further are called the terminal nodes. Considering a certain split point in the values of one of the explanatory variables, each nonterminal node is split into two descendant nodes. Values of the covariate that are below this split point go to the left descendant node, while higher values go to the right one.

To determine the exact split point to partition a node, every possible split in all covariates considered in the tree is taken into account. In regression trees, the splitting criterion is based on the mean squared error of the predictions of the descendant nodes. Hence, the squared error of the prediction in a node can be written as:

$$Q = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (4)$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the prediction of y at the given node.

Since each split leads to two descendant nodes, the Qs of both descendant nodes as well as the sample sizes of both nodes are considered in the choice of the optimal split point. It is determined by minimizing:

$$Q_{split} = n_L Q_L + n_R Q_R \quad (5)$$

where Q_L and Q_R are the mean squared errors in the predictions of the left and right descendant node and n_L and n_R are the corresponding sample sizes.

After the selection of a split point, the node is partitioned into its two descendant nodes, which are again split in the same way. This procedure is continued until either a predefined level of Q_{split} or a predefined maximal tree size is reached.

Each of the trees can then be used to obtain an individual estimate of the response variable y. Finally, the simple average of the estimations of all trees gives the actual Random Forest prediction.

(2) Tuning

When estimating a Random Forest, the parameter that the estimation can be most sensitive to is the value m of randomly selected covariates chosen at each node of the trees. This parameter can be tuned and optimized using out-of-bag error estimates.

Since with bootstrapping only a random subset of the training data is used in each tree, there is always some training data left that is not employed in the individual estimations. These parts are called out-of-bag data and can be used to compute out-of-bag error rates. Tuning the parameter m entails calculating out-of-bag error rates for different values of m and then choosing m such that the out-of-bag error is minimized.

Another parameter that can potentially be tuned is the value J of trees in the forest. However, Breiman (2001) has showed that as J increases, the out-of-bag error converges to a limit. Hence, it is sufficient to ensure that J is not set too low, but no actual tuning is needed for this part of a Random Forest estimation.

B. Bayesian Model Averaging

Since there is limited research on how to produce reasonable forecasts of population density on a granular level, we want to consider different approaches for this task. The first one is a Random Forest technique, which we also used to obtain grid-level information on population density for years where we have government-published population numbers. The second one is Bayesian model averaging (BMA).

BMA is an ensemble learning method that uses Bayesian inference to solve the problem of model selection. It considers a large number of models, evaluates their explanatory power, and then weights the variables accordingly. An advantage of this approach is that it facilitates the inclusion of a large set of covariates and internally weights them by their relative importance in explaining the dependent variable. That way, it considers the uncertainty regarding which explanatory factors to include. Since we know that the estimates of coefficients depend on the covariates entering the model and there is not much literature on which covariates to employ in our forecasts, model uncertainty has to be taken into account. Hence, BMA could be a suitable approach.

As opposed to Random Forests, BMA does not create regression trees, but relies on simple linear regressions.[3] Consequently it does not search for split points in the data, but rather directly estimates the average effect each covariate has on the dependent variable. Furthermore, BMA does not use random subsets of the explanatory variables but evaluates all different combinations of covariates that can be created from the input data. Another difference is that BMA does not use subsets of the observations in the individual regressions, but rather uses all observations in each of its regressions. Finally, while Random Forests simply use the average of the predictions of all individual trees as their result, BMA uses weighted averages, giving more weight to individual models with higher explanatory power.

To apply this method on the estimation of the population growth rate, we set up a regression of the following form:

$$y = X_k \beta_k + \varepsilon \tag{6}$$

In this linear regression, y is a column vector of population growth rates. X_k is a matrix of k columns, one for each explanatory variable, and β_k is a k-dimensional vector of parameters corresponding to the variables in X_k.

As explanatory variables, we will employ all covariates used to estimate grid-level population density in past years as well as the initial population density. Details on all data sets entering the estimation can be found in section 2 of this paper. Regarding the initial population density, we tried two different ways to include this variable. First, we included it in the form of absolute numbers of people. Second, we conducted the same estimations with these levels replaced by logs of population density.

To apply BMA on the linear regression given above, we estimated sets of linear regression models. This means that we considered all models that can be set up as combinations of the variables included in X_k. Since X_k consists of k potential covariates, we have a total of 2^k possible variable combinations and a set of potential models $M = \{M_1, M_2, \ldots, M_{2^k-1}, M_{2^k}\}$. The BMA estimation assesses all these models by their ability to explain the dependent variable and then weights the models by their explanatory power. As a result of this evaluation, weighted averages of the models are used to carry out inference.

To produce these weighted averages of all models, BMA uses weights that are based on the posterior model probabilities resulting from Bayes' theorem. Thus, the posterior model probability of a model M_γ can be written as follows:

$$p(M_\gamma | y, X) = \frac{p(y|M_\gamma, X) p(M_\gamma)}{p(y|X)} = \frac{p(y|M_\gamma, X) p(M_\gamma)}{\sum_{s=1}^{2^K} p(y|M_s, X) p(M_s)} \tag{7}$$

In this equation, the marginal likelihood $p(M_\gamma)$ refers to the prior model probability, expressing the researcher's opinion of the likelihood of model M_γ before checking the data. To get to the posterior model probability, it is updated by the information in the data. This is done by multiplying $p(M_\gamma)$ with the

[3] Since there is limited literature that suggests specific types of nonlinearities to consider in this context, we started with a simple linear model to get an idea whether BMA could be a suitable alternative to random forest. Future studies could further explore different specifications of BMA to make a conclusive assessment on whether random forest still performs better.

integrated likelihood $p(y|M_\gamma, X)$, representing the probability of observing the given data based on the model M_γ. $p(y|M_\gamma, X)$ denotes the likelihood of y integrated over the parameter space for a given model M_γ. In contrast, $p(y|X)$ expresses the likelihood of y integrated over the model space and does not depend on a specific model. Hence, it remains a constant multiplicative term (Zeugner and Feldkircher 2015).

Given the posterior model probabilities of all models included in M, it is now possible to derive the marginal posterior distribution of any statistic present in the models. For example, the model-weighted posterior distribution of the estimator of the coefficient β_k can be described as:

$$p(\beta_k|y, X) = \sum_{\gamma=1}^{2^k} p(\beta_k|y, X, M_\gamma)\, p(M_\gamma|y, X) \tag{8}$$

This means that the posterior distribution of the estimator of β_k given y and X can be written as the average of its posterior distribution under each of the models in M, weighted by their posterior model probabilities (Hoeting et al. 1999; Fragoso, Bertoli, and Louzada 2018).

(1) Priors

To obtain posterior model probabilities and posterior distributions of the model parameters, corresponding priors have to be chosen. If knowledge from previous studies is available, these priors should reflect the researchers' prior beliefs. However, like in the case of this study, prior information is often limited. In such cases, using a uniform prior model probability is a common choice (Zeugner and Feldkircher 2015).

Regarding prior beliefs about the coefficients, the common practice is to assume a normal distribution and specify the expected mean and standard deviation (Zeugner and Feldkircher 2015). In particular, the prior mean is usually set to zero and the variance is based on Zellner's g:

$$g\sigma^2 (X_k{}^{\mathrm{T}}\, X_k)^{-1} \tag{9}$$

Hence, the prior distribution of β_k can be written as:

$$\beta_k|g \sim N\left(0, g\sigma^2 \left(X_k{}^{\mathrm{T}}\, X_k\right)^{-1}\right) \tag{10}$$

This specification reflects that the researchers expect the coefficient to be zero and its variance to be closely related to that of the explanatory variables in X_k. As the equation shows, the hyperparameter g defines how large the variance of β_k is expected to be. For this paper g is set to:

$$g = \max\,(n, k^2) \tag{11}$$

where n is the number of observations and k is the number of covariates. This approach follows the work of Fernández, Ley, and Steel (2001).

(2) Model sampling

Once these priors are set, BMA is applied to estimate the posterior model probabilities as well as the posterior distributions of the coefficients. The high cardinality of the model space given in this and many other applications of BMA often makes the specific evaluation of every single potential model infeasible. As one solution to this challenge, Markov chain Monte Carlo model composition (MC^3) has proven to conduct sensible evaluations of subsets of the model space that account for a reasonably large part of the posterior model probability mass (Crespo Cuaresma and Feldkircher 2013). This motivated us to evaluate the set of potential models with a random walk MC^3 search algorithm.

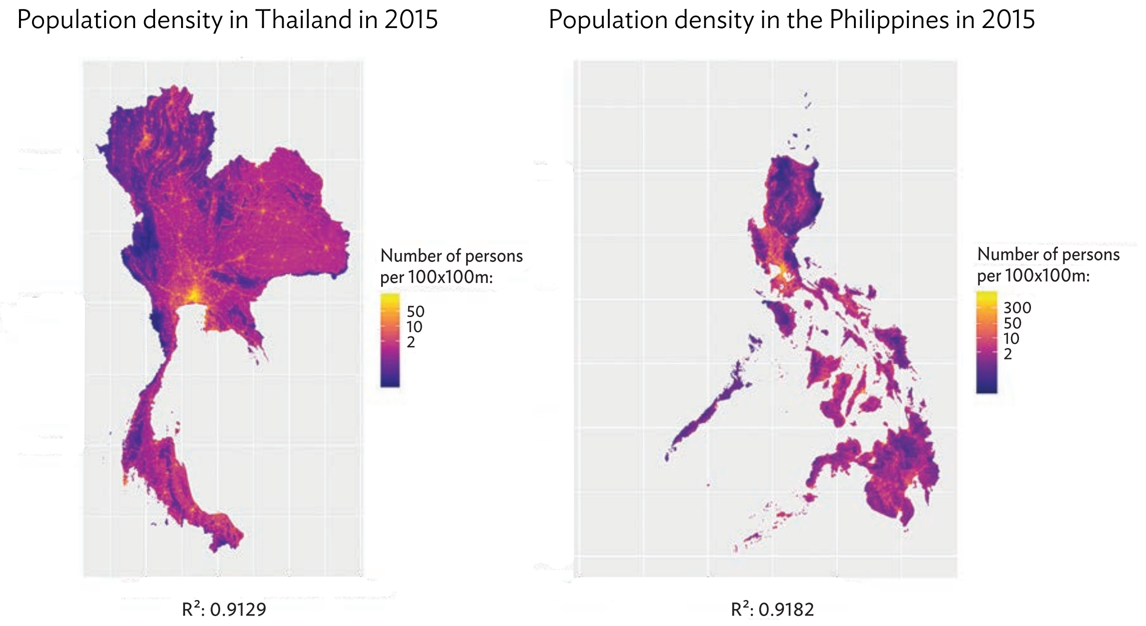
IV. RESULTS

To obtain grid-level estimates of population density for Thailand and the Philippines, we first overlaid all input data for the years for which we have government-published population data. For Thailand, we have access to *tambon*-level data for the years 2013, 2015, and 2017. For the Philippines, we have municipal-level data for 2012, 2015, and 2018. Next, we applied a Random Forest approach as described in Section III with the log of population density as the dependent variable and parameters like land cover classes, elevation, slope, and nighttime lights as covariates, using data aggregated at the *barangay* and *tambon* level. The resulting (adjusted) R^2 value is 0.92 for both countries. By applying the trained model to the grid-level data of the explanatory variables, we are able to estimate population density on the 100 meter by 100 meter grid level for all years for which we have input data from government-published estimates.

In particular, the model predictions are used as weights to distribute the population inside a *tambon* or municipality at the grid level. This means that we first create estimates of the total population in each 100 meter by 100 meter grid. Next, we rescale these numbers to make them sum up to our input data on the *tambon* or municipality level. This ensures that our results always match the official data on the most granular administrative level included in the census data.

Figure 3 shows the results of the Random Forest predictions of population density on the grid level for Thailand and the Philippines for the year 2015.

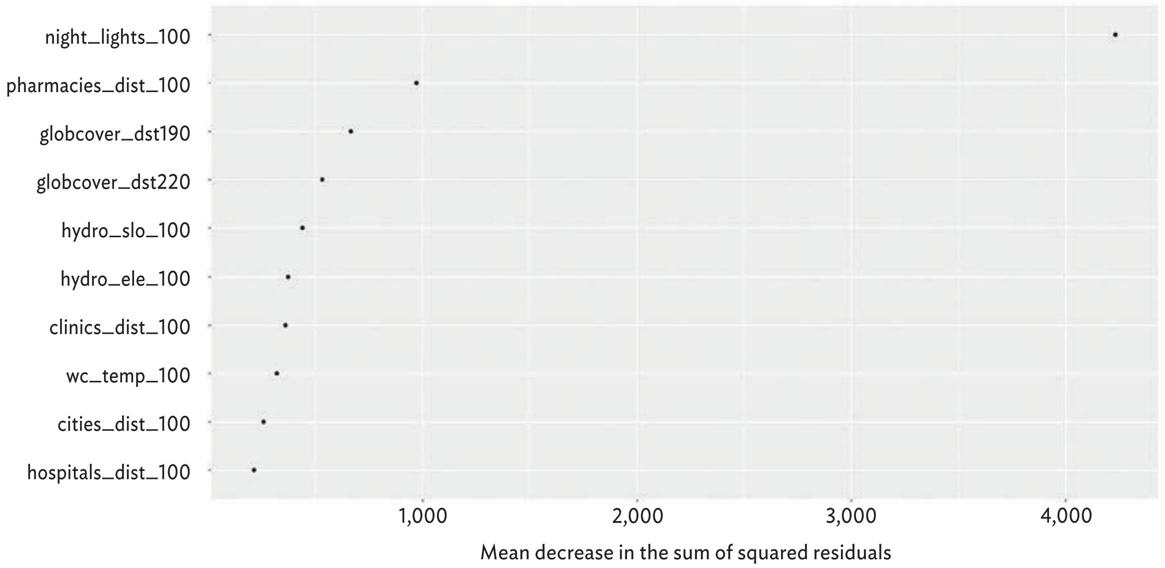
Figure 3: Random Forest Prediction of Population Density in 2015


Population density in Thailand in 2015
Population density in the Philippines in 2015
Number of persons per 100x100m:
50
10
2
Number of persons per 100x100m:
300
50
10
2
R^2: 0.9129
R^2: 0.9182

Source: Authors' calculations using data from the National Statistical Office of Thailand, Philippine Statistics Authority, GlobCover of the European Space Agency, HydroSHEDS of the Conservation Science Program of World Wildlife Fund, Moderate Resolution Imaging Spectroradiometer and Visible Infrared Imaging Radiometer Suite of the National Aeronautics and Space Administration, WorldClim, OpenStreetMap, and Protected Planet.

Apart from these predictions, the model also shows the relative importance of different independent variables in explaining the dependent variable. Figure 4 gives the explanatory power of the 10 most relevant covariates in describing population density on the grid level in Thailand in 2015. Details on all variables can be found in Table 1 in Section II.

Figure 4: Random Forest Variable Importance of Top 10 Covariates in Thailand in 2015



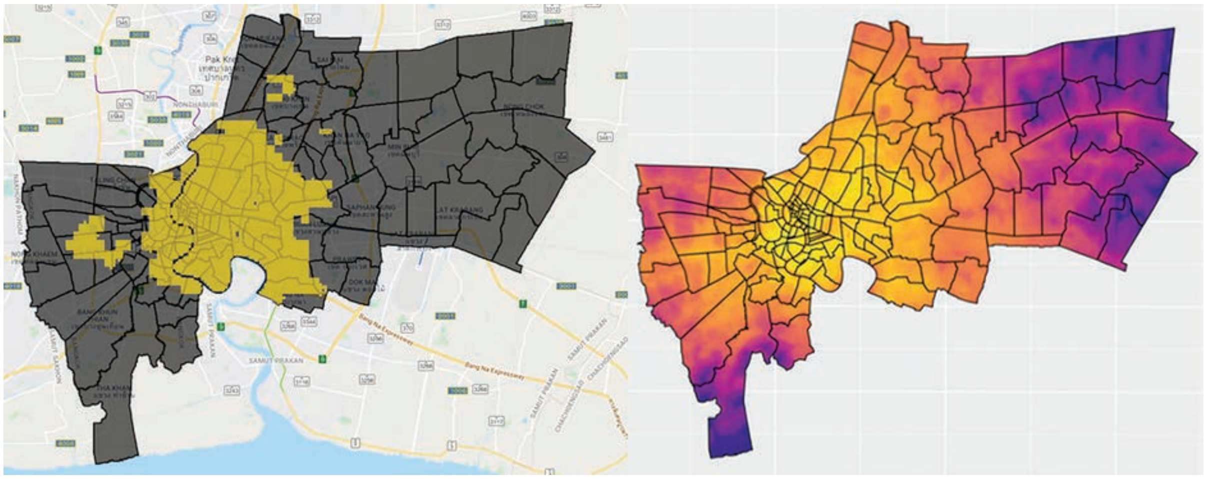
Notes: Presented as the mean decrease in the sum of squared residuals when the variable is included in a tree split. For the description of the variables in the vertical axis, please see Appendix.

Source: Authors' calculations using data from the National Statistical Office of Thailand, GlobCover of the European Space Agency, HydroSHEDS of the Conservation Science Program of World Wildlife Fund, Moderate Resolution Imaging Spectroradiometer and Visible Infrared Imaging Radiometer Suite of the National Aeronautics and Space Administration, WorldClim, OpenStreetMap, and Protected Planet.

As Figure 4 shows, one of the most important explanatory factors is the variable globcover_dst190, which captures the distance to the next grid covered by the land cover class 190, representing artificial surfaces and associated areas. To get a better picture of how this factor is related to population density, Figure 5 illustrates the areas covered by this land cover class next to our results on the population density in Bangkok in 2015.

Comparing the most important variables across our estimations for different years, we find very similar patterns. Regarding a comparison of the most important variables for Thailand versus the Philippines, we can see that while most variables are still the same, we can notice some differences. Like in Thailand, data on lights at night is also the most important predictor of population density in the Philippines. Similarly, the distances to pharmacies and hospitals as well as data on temperature and artificial surfaces also belong to the top ten of the most important variables for estimating population density in the Philippines. However, the main difference is that slope and elevation seem to be less important than in Thailand. This is probably related to the differences in the geographies of the two

Figure 5: Land Cover Class 190 versus Population Density in Bangkok in 2015



Source: Authors' calculations using data from the National Statistical Office of Thailand, GlobCover of the European Space Agency, HydroSHEDS of the Conservation Science Program of World Wildlife Fund, Moderate Resolution Imaging Spectroradiometer and Visible Infrared Imaging Radiometer Suite of the National Aeronautics and Space Administration, WorldClim, OpenStreetMap, and Protected Planet.

countries. Except for its mountains in the north, large parts of Thailand are very flat, while due to their volcanic origin the islands of the Philippines are generally more mountainous.[4]

Building on the results for Thailand for the years 2013, 2015, and 2017, we extend our estimations by forecasting population density for three provinces for the year 2020. Here we use population growth rate as the dependent variable and the initial level of population density as well as the remotely sensed input data from our earlier predictions as explanatory factors.

Facing greater uncertainty regarding how to produce reasonable forecasts of population density on this level of granularity, we aim at assessing different methods and model specifications. In particular, we considered a Random Forest and a BMA approach. Additionally, we evaluated how to include the initial level of population density in our model, either in absolute terms or in logs. The evaluation of the suitability of these different approaches is based on data from 2013 and 2015, which are used to predict population density in Thailand in 2017. These numbers are then compared to actual data from the same year by computing the root mean squared error and the mean average error of our predictions. Table 1 shows the results of this comparison.

[4] Comparing these results with that of Stevens et al. (2015), we find similarities and differences. For example, like in our estimations, lights at night play a very important role and are among the most central covariates. Similarly, the distances to the next urban area and to the next hospital also seem to play an important role in some countries considered in Stevens et al. (2015). In addition, the distances to the next generic health facility as well as to the settlement point and the next community are also highly relevant for some of the models in Stevens et al. (2015).

Table 1: Root Mean Square Error and Mean Average Error of Forecasts of Population Density for Thailand in 2017

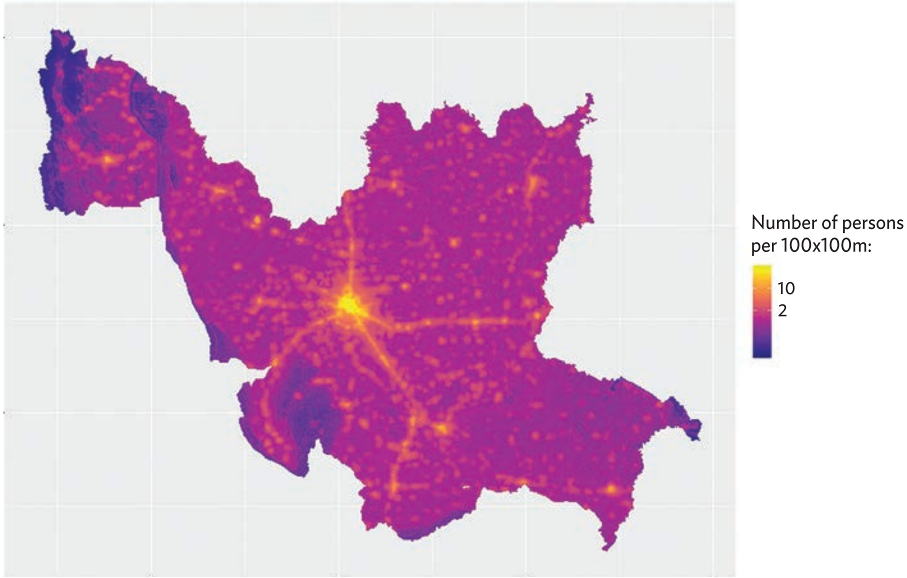
	Root Mean Squared Error		Mean Average Error	
	Random Forest	**Bayesian Model Averaging**	**Random Forest**	**Bayesian Model Averaging**
Dependent variable: natural logarithm of population	3.27	56.13	1.81	27.68
Dependent variable: population in absolute numbers	3.27	52.02	1.79	27.08

Source: Authors' calculations using data from the National Statistical Office of Thailand, GlobCover of the European Space Agency, HydroSHEDS of the Conservation Science Program of World Wildlife Fund, Moderate Resolution Imaging Spectroradiometer and Visible Infrared Imaging Radiometer Suite of the National Aeronautics and Space Administration, WorldClim, OpenStreetMap, and Protected Planet.

As Table 1 shows, the Random Forest approach clearly outperforms BMA in our exercise. Regarding the specification of the initial population, the root mean square error is slightly lower with population in logs, while the mean average error is slightly lower with population in absolute numbers. Hence, we use a Random Forest to conduct our actual forecasts and we use the log of the initial population in our set of covariates.

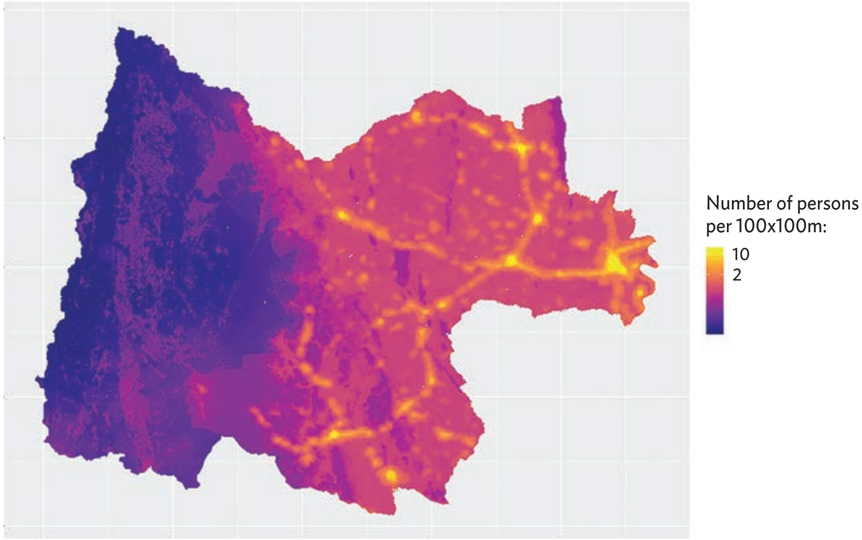
To obtain grid-level forecasts of population density in the provinces of Udon Thani, Uthai Thani, and Samut Songkhram for the year 2020, we first trained the model with information on the average annual population growth in Thailand from 2013 to 2017 and initial data from 2013. Second, we use the trained Random Forest and initial data from 2017 to estimate annual growth rates after 2017. Applying these growth rates to grid-level population data from 2017, we finally obtained granular forecasts of population density in 2020. Figures 6 to 8 show the results of our forecasts of population density on the 100 meters by 100 meters grid level for Udon Thani, Uthai Thani, and Samut Songkhram.

Figure 6: Forecasts of Population Density in Udon Thani in 2020



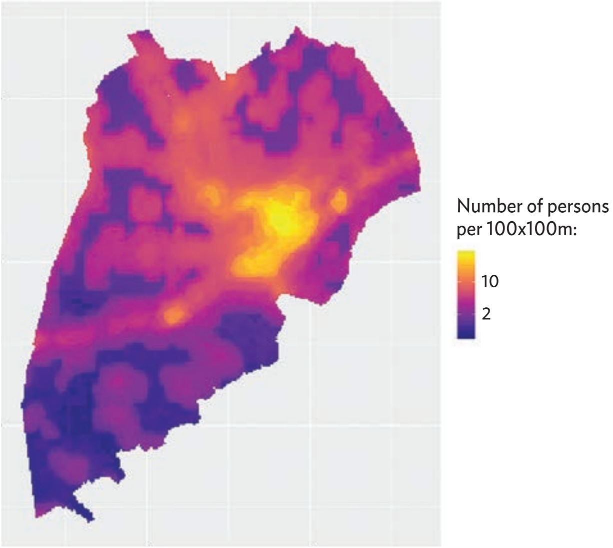
Source: Authors' calculations using data from the National Statistical Office of Thailand, GlobCover of the European Space Agency, HydroSHEDS of the Conservation Science Program of World Wildlife Fund, Moderate Resolution Imaging Spectroradiometer and Visible Infrared Imaging Radiometer Suite of the National Aeronautics and Space Administration, WorldClim, OpenStreetMap, and Protected Planet.

Figure 7: Forecasts of Population Density in Uthai Thani in 2020



Source: Authors' calculations using data from the National Statistical Office of Thailand, GlobCover of the European Space Agency, HydroSHEDS of the Conservation Science Program of World Wildlife Fund, Moderate Resolution Imaging Spectroradiometer and Visible Infrared Imaging Radiometer Suite of the National Aeronautics and Space Administration, WorldClim, OpenStreetMap, and Protected Planet.

Figure 8: Forecasts of Population Density in Samut Songkhram in 2020

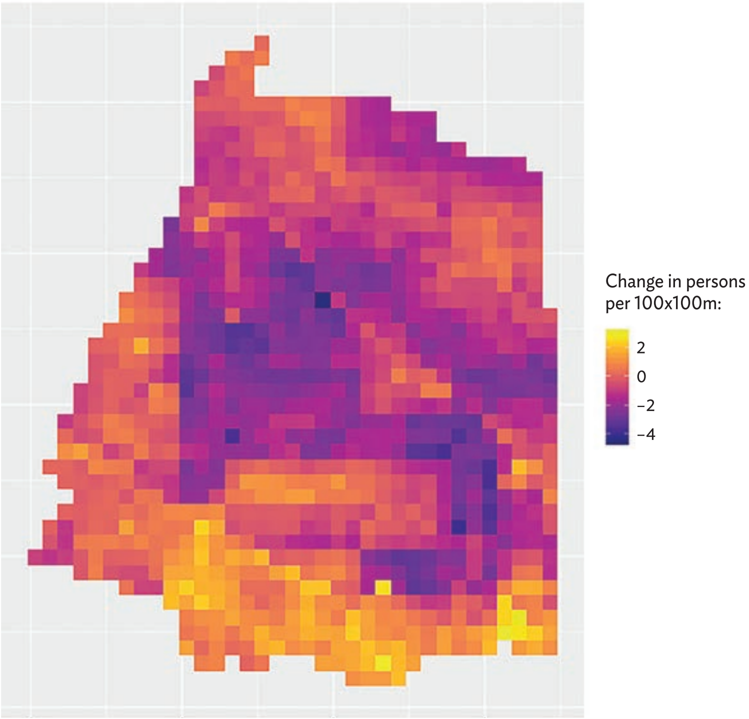


Source: Authors' calculations using data from the National Statistical Office of Thailand, GlobCover of the European Space Agency, HydroSHEDS of the Conservation Science Program of World Wildlife Fund, Moderate Resolution Imaging Spectroradiometer and Visible Infrared Imaging Radiometer Suite of the National Aeronautics and Space Administration, WorldClim, OpenStreetMap, and Protected Planet.

These forecasts allow us to project the change in population density in these provinces from 2013 to 2020. Focusing on the center of the city of Udon Thani, the capital of the eponymous province, we can see that population density has decreased in some areas, while it has increased in other parts of the city. Figure 9 shows the expected change in the population density of the Tambon Mak Khaeng in the city of Udon Thani from 2013 to 2020 on the 100 meter by 100 meter grid level.[5]

[5] Since estimates for 2020 are not rescaled, we are allowing for the total population to grow inside the areas under consideration. For each grid, we predict the growth in population density from 2017 to 2020 and then apply these individual growth rates to the results for 2017. This does not only change the population density on the grid level, but also leads to a change in the projected total population in each of the districts. However, for long-term forecasts, including detailed information on migration dynamics could potentially improve the forecasts further.

Figure 9: Expected Change in Population Density in Mak Khaeng, Udon Thani from 2013 to 2020



Source: Authors' calculations using data from the National Statistical Office of Thailand, GlobCover of the European Space Agency, HydroSHEDS data of the Conservation Science Program of World Wildlife Fund, Moderate Resolution Imaging Spectroradiometer and Visible Infrared Imaging Radiometer Suite of the National Aeronautics and Space Administration, WorldClim, OpenStreetMap, and Protected Planet.

V. SUMMARY AND CONCLUSION

Availability of accurate, timely, and spatially disaggregated distributions of human population are important on many fronts. Population data are used as inputs when national governments allocate resources across its local territories. Businesses and other nongovernment institutions also use population data to strategize their operations. At the global level, population data are also fundamental in monitoring the Sustainable Development Goals as many of its indicators require population data as baseline information.

In developing countries, particularly in Asia and the Pacific, population and household census is conventionally used to compile population numbers. Despite being considered as a useful and reliable collection vehicle for population data, it suffers from several limitations. Censuses are costly and hence, these are only conducted every 5 to 10 years in many countries. There is also a need for more spatially granular population estimates than data conventionally published from population and household censuses as decision-making operates at finer levels.

This paper uses a combination of published population data (mostly derived from census information) and publicly available data based on satellite imagery to produce estimates of population density on the 100 meters by 100 meters grid level. Focusing on Thailand and the Philippines, we follow Stevens et al. (2015) and apply a Random Forest approach. In addition to predictions for past years for which published population data are available, we also explored a population forecasting method on the same level of granularity in select provinces of Thailand. For this part of our estimations, we evaluate different model specifications and, in addition to Random Forests, also consider a BMA approach. Our results show that using a Random Forest model with the log of the initial population and remotely sensed data as covariates, reasonable forecasts of grid-level population growth rates are achievable. This method allows us produce forecasts of population density for three provinces in Thailand for the year 2020.

The results of this paper contribute to the assessment of ensemble methods like Random Forest and BMA in the field of demography. They showcase the applicability of these methods, and especially a Random Forest approach, in grid-level predictions of population density and might be a starting point for further and more extensive forecasts. Especially in times of continued population growth and urbanization, this area of research has the potential to contribute to solving challenges in a variety of fields like economics, health, and environmental policy.

APPENDIX

Table A1: Description of Variables Used in Estimation of Population Density

Type	Variable Name(s)	Description	Thailand	Philippines
Census	y_data	Country-specific census and scale	National census, Tambon level	National census, Municipality level
Land Cover	globcover_cls11/ globcover_dst11[1]	Post-flooding or irrigated croplands (or aquatic)	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls14/ globcover_dst14	Rainfed croplands	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls20/ globcover_dst20	Mosaic cropland (50%–70%)/ vegetation (grassland/shrubland/ forest) (20%–50%)	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls30/ globcover_dst30	Mosaic vegetation (grassland/shrubland/ forest) (50%–70%)/ cropland (20%–50%)	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls40/ globcover_dst40	Closed to open (>15%) broadleaved evergreen or semi-deciduous forest (>5m)	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls50/ globcover_dst50	Closed (>40%) broadleaved deciduous forest (>5m)	GlobCover, 300m	
Land Cover	globcover_cls60/ globcover_dst60	Open (15%–40%) broadleaved deciduous forest/woodland (>5m)	GlobCover, 300m	
Land Cover	globcover_cls70/ globcover_dst70	Closed (>40%) needleleaved evergreen forest (>5m)	GlobCover, 300m	
Land Cover	globcover_cls100/ globcover_dst100	Closed to open (>15%) mixed broadleaved and needleleaved forest (>5m)	GlobCover, 300m	
Land Cover	globcover_cls110/ globcover_dst110	Mosaic forest or shrubland (50%–70%)/grassland (20%–50%)	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls120/ globcover_dst120	Mosaic grassland (50%–70%)/ forest or shrubland (20%–50%)	GlobCover, 300m	

continued on next page

Table A1 *continued*

Type	Variable Name(s)	Description	Thailand	Philippines
Land Cover	globcover_cls130/ globcover_dst130	Closed to open (>15%) (broadleaved or needleleaved, evergreen or deciduous) shrubland (<5m)	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls140/ globcover_dst140	Closed to open (>15%) herbaceous vegetation (grassland, savannas or lichens/mosses)	GlobCover, 300m	
Land Cover	globcover_cls150/ globcover_dst150	Sparse (<15%) vegetation	GlobCover, 300m	
Land Cover	globcover_cls160/ globcover_dst160	Closed to open (>15%) broadleaved forest regularly flooded (semi-permanently or temporarily)- Fresh or brackish water	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls170/ globcover_dst170	Closed (>40%) broadleaved forest or shrubland permanently flooded- Saline or brackish water	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls180/ globcover_dst180	Closed to open (>15%) vegetation (grassland, shrubland, woody vegetation) on regularly flooded or waterlogged soil-fresh, brackish or saline water	GlobCover, 300m	
Land Cover	globcover_cls190/ globcover_dst190	Artificial surfaces and associated areas (Urban areas >50%)	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls200/ globcover_dst200	Bare areas	GlobCover, 300m	
Land Cover	globcover_cls210/ globcover_dst210	Water bodies	GlobCover, 300m	GlobCover, 300m
Land Cover	globcover_cls220/ globcover_dst220	Permanent snow and ice	GlobCover, 300m	GlobCover, 300m
Protected Areas	protected_areas_100/ protected_areas_dist _100[2]	Protected area	Protected Planet	Protected Planet
Map Features	cities_100/ cities_dist_100	City	OpenStreetMap	OpenStreetMap
Map Features	clinics_100/ clinics_dist_100	Clinic	OpenStreetMap	OpenStreetMap
Map Features	hamlets_100/ hamlets_dist_100	Hamlet	OpenStreetMap	OpenStreetMap

continued on next page

Table A1 *continued*

Type	Variable Name(s)	Description	Thailand	Philippines
Map Features	hospitals_100/ hospitals_dist_100	Hospital	OpenStreetMap	OpenStreetMap
Map Features	pharmacies_100/ pharmacies_dist_100	Pharmacy	OpenStreetMap	OpenStreetMap
Map Features	railways_100/ railways_dist_100	Railway	OpenStreetMap	OpenStreetMap
Map Features	rivers_100/ rivers_dist_100	River	OpenStreetMap	OpenStreetMap
Map Features	schools_100/ schools_dist_100	School	OpenStreetMap	OpenStreetMap
Map Features	suburbs_100/ suburbs_dist_100	Suburb	OpenStreetMap	OpenStreetMap
Map Features	towns_100/ towns_dist_100	Town	OpenStreetMap	OpenStreetMap
Map Features	villages_100/ villages_dist_100	Village	OpenStreetMap	OpenStreetMap
Map Features	water_100/ water_dist_100	Water	OpenStreetMap	OpenStreetMap
Elevation	hydro_ele_100	Elevation	HydroSHEDS, 100m	HydroSHEDS, 100m
Slope	hydro_slo_100	Slope	HydroSHEDS, 100m	HydroSHEDS, 100m
Net Primary Production	modis_100	Amount of carbon captured by plants	MODIS, 250m	MODIS, 250m
Precipitation	wc_prec_100	Monthly data on precipitation	WorldClim, 1km	WorldClim, 1km
Temperature	wc_temp_100	Monthly data on temperature	WorldClim, 1km	WorldClim, 1km
Nighttime Lights	night_lights_100	Lights at night	VIIRS, 500m	VIIRS, 500m

km = kilometer, m = meter.

Source: Authors’ compilation.

REFERENCES

Anderson, Delmar, and Philip Anderson. 1973. "Population Estimates by Humans and Machines." *Photogrammetric Engineering* 39 (2): 147–54.

Azar, Derek, Ryan Engstrom, Jordan Graesser, and Joshua Comenetz. 2013. "Generation of Fine-Scale Population Layers Using Multi-Resolution Satellite Imagery and Geospatial Data." *Remote Sensing of Environment* 130: 219–32. doi: https://doi.org/10.1016/j.rse.2012.11.022.

Azar, Derek, Jordan Graesser, Ryan Engstrom, Joshua Comenetz, Robert M. Leddy, Nancy G. Schechtman, and Theresa Andrews. 2010. "Spatial Refinement of Census Population Distribution Using Remotely Sensed Estimates of Impervious Surfaces in Haiti." *International Journal of Remote Sensing* 31 (21): 5635–55. doi: 10.1080/01431161.2010.496799.

Balk, Deborah L., Uwe Deichmann, Greg Yetman, Francesca Pozzi, Simon I. Hay, and Andrew Nelson. 2006. "Determining Global Population Distribution: Methods, Applications and Data." *Advances in Parasitology* 62: 119–56. doi: 10.1016/s0065-308x(05)62004-0.

Balk, Deborah, and Greg Yetman. 2004. "The Global Distribution of Population: Evaluating the Gains in Resolution Refinement." Center for International Earth Science Information Network, Columbia University, Palisades, NY.

Bhaduri, Budhendra, Edward Bright, Phillip Coleman, and Marie L. Urban. 2007. "Landscan USA: A High-Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics." *GeoJournal* 69 (1): 103–17. doi: 10.1007/s10708-007-9105-9.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5-32. doi: 10.1023/A:1010933404324.

Chen, K. 2002. "An Approach to Linking Remotely Sensed Data and Areal Census Data." *International Journal of Remote Sensing* 23 (1): 37–48. doi: 10.1080/01431160010014297.

Crespo Cuaresma, Jesús, and Martin Feldkircher. 2013. "Spatial Filtering, Model Uncertainty and the Speed of Income Convergence in Europe." *Journal of Applied Econometrics* 28 (4): 720–41. doi: https://doi.org/10.1002/jae.2277.

Cutler, Adele, D. Richard Cutler, and John R. Stevens. 2012. "Random Forests." In *Ensemble Machine Learning: Methods and Applications*, edited by Cha Zhang and Yunqian Ma, 157–75. Boston, MA: Springer.

Deichmann, Uwe. 1996. "A Review of Spatial Population Database Design and Modeling." NCGIA Technical Report 96-3. National Center for Geographic Information and Analysis, Department of Geography, University of California, Santa Barbara, CA.

Dobson, Jerome E., Edward A. Brlght, Phllllp R. Coleman, Rlchard C. Durfee, and Brian A. Worley. 2000. "Landscan: A Global Population Database for Estimating Populations at Risk." *Photogrammetric Engineering and Remote Sensing* 66 (7): 849–57.

Fernández, Carmen, Eduardo Ley, and Mark F. J. Steel. 2001. "Benchmark Priors for Bayesian Model Averaging." *Journal of Econometrics* 100 (2): 381–427. doi: https://doi.org/10.1016/S0304-4076(00)00076-2.

Fragoso, Tiago M., Wesley Bertoli, and Francisco Louzada. 2018. "Bayesian Model Averaging: A Systematic Review and Conceptual Classification." *International Statistical Review* 86 (1): 1–28. doi: https://doi.org/10.1111/insr.12243.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14 (4): 382–401.

Linard, Catherine, Marius Gilbert, and Andrew J. Tatem. 2011. "Assessing the Use of Global Land Cover Data for Guiding Large Area Population Distribution Modelling." *GeoJournal* 76 (5): 525–38. doi: 10.1007/s10708-010-9364-8.

Linard, Catherine, Marius Gilbert, Robert W. Snow, Abdisalan M. Noor, and Andrew J. Tatem. 2012. "Population Distribution, Settlement Patterns and Accessibility across Africa in 2010." *PLOS ONE* 7 (2): e31743. doi: 10.1371/journal.pone.0031743.

Lung, Tobias, Tillmann Lübker, James K. Ngochoch, and Gertrud Schaab. 2013. "Human Population Distribution Modelling at Regional Level Using Very High Resolution Satellite Imagery." *Applied Geography* 41: 36–45. doi: https://doi.org/10.1016/j.apgeog.2013.03.002.

Salvatore, Mirella, Francesca Pozzi, Ergin Ataman, Barbara Huddleston, and Mario Bloise. 2005. "Mapping Global Urban and Rural Population Distributions." Environment and Natural Resources Working Paper No. 24. Food and Agriculture Organization of the United Nations, Rome.

SDSN. 2015. *Data for Development: A Needs Assessment for SDG Monitoring and Statistical Capacity Development*. New York, NY: Sustainable Development Solutions Network.

Stevens, Forrest R., Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. 2015. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data." *PLOS ONE* 10 (2): e0107042. doi: 10.1371/journal.pone.0107042.

Sutton, P., D. Roberts, C. Elvidge, and K. Baugh. 2001. "Census from Heaven: An Estimate of the Global Human Population Using Night-Time Satellite Imagery." *International Journal of Remote Sensing* 22 (16): 3061–76. doi: 10.1080/01431160010007015.

Tatem, Andrew J., Susana Adamo, Nita Bharti, Clara R. Burgert, Marcia Castro, Audrey Dorelien, Gunter Fink, Catherine Linard, Mendelsohn John, Livia Montana, Mark R. Montgomery, Andrew Nelson, Abdisalan M. Noor, Deepa Pindolia, Greg Yetman, and Deborah Balk. 2012. "Mapping Populations at Risk: Improving Spatial Demographic Data for Infectious Disease Modeling and Metric Derivation." *Population Health Metrics* 10 (1): 8. doi: 10.1186/1478-7954-10-8.

Tatem, Andrew J., Abdisalan M. Noor, Craig von Hagen, Antonio Di Gregorio, and Simon I. Hay. 2007. "High Resolution Population Maps for Low Income Nations: Combining Land Cover and Census in East Africa." *PLOS ONE* 2 (12): e1298. doi: 10.1371/journal.pone.0001298.

Tobler, Waldo, Uwe Deichmann, Jon Gottsegen, and Kelly Maloy. 1995. "The Global Demography Project." Technical Report TR-95-6. National Center for Geographic Information and Analysis, Department of Geography, University of California, Santa Barbara, CA.

UNCTAD. 2021. COVID-19: Measurement Issues and Assessments. SDG Pulse. July 2. https://sdgpulse.unctad.org/covid-19/.

UN DESA. 2018. *2018 Revision of World Urbanization Prospects*. New York, NY: United Nations Department of Economic and Social Affairs.

Zeugner, Stefan, and Martin Feldkircher. 2015. "Bayesian Model Averaging Employing Fixed and Flexible Priors: The BMS Package for R." *Journal of Statistical Software* 68 (4): 1–37. doi: 10.18637/jss.v068.i04.

Compiling Granular Population Data Using Geospatial Information

This paper demonstrates ensemble methods in population mapping and assesses their applicability in analyzing and forecasting population density on a grid level. The study uses a Random Forest approach to estimate population density in the Philippines and Thailand on a 100-meter by 100-meter grid level. The study also uses different specifications of Random Forest and Bayesian model averaging techniques to create grid-level population density forecasts in three provinces of Thailand.

About the Asian Development Bank

ADB

ASIAN DEVELOPMENT BANK
6 ADB Avenue, Mandaluyong City
1550 Metro Manila, Philippines
www.adb.org