Policy Research Working Paper 8776

Estimating Small Area Population Density Using Survey Data and Satellite Imagery

An Application to Sri Lanka

Ryan Engstrom
David Newhouse
Vidhya Soundararajan

WORLD BANK GROUP
Poverty and Equity Global Practice
March 2019

Policy Research Working Paper 8776

Abstract

Country-level census data are typically collected once every 10 years. However, conflict, migration, urbanization, and natural disasters can cause rapid shifts in local population patterns. This study uses Sri Lankan data to demonstrate the feasibility of a bottom-up method that combines household survey data with contemporaneous satellite imagery to track frequent changes in local population density. A Poisson regression model based on indicators derived from satellite data, selected using the least absolute shrinkage and selection operator, accurately predicts village-level population density. The model is estimated in villages sampled in the 2012/13 Household Income and Expenditure Survey to obtain out-of-sample density predictions in the nonsurveyed villages. The predictions approximate the 2012 census density well and are more accurate than other bottom-up studies based on lower-resolution satellite data. The predictions are also more accurate than most publicly available population products, which rely on areal interpolation of census data to redistribute population at the local level. The accuracies are similar when estimated using a random forest model, and when density estimates are expressed in terms of population counts. The collective evidence suggests that combining surveys with satellite data is a cost-effective method to track local population changes at more frequent intervals.

This paper is a product of the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/research. The authors may be contacted at vidhyasrajan@iimb.ac.in.

Estimating Small Area Population Density Using Survey Data and Satellite Imagery: An Application to Sri Lanka*

Ryan Engstrom† David Newhouse‡ Vidhya Soundararajan§

Keywords: Population density predictions, Satellite imagery, Machine learning.

JEL Code: J19, C52, C53.

* We thank Sarah Antos for help with the data, and Jonathan Hersh for comments and discussions, and presenting an earlier version of this paper. We also thank Sarosh Sattar, Ani Silwal, and Walker Bradley for detailed comments, and the participants of Geo4dev symposium 2017, Population Association of America Annual conference 2018, and seminar participants of the World Bank for comments and suggestions. We acknowledge the funding for this project from the World Bank group. All remaining errors are our own.

† George Washington University, Washington, DC. Email: rengstro@gwu.edu

‡ The World Bank, Washington DC. Email: dnewhouse@worldbank.org

§ Indian Institute of Management Bangalore. Email: vidhyasrajan@iimb.ernet.in

1 Introduction

Up-to-date estimates of population density in small areas is a valuable input for policymakers (Stevens et al., 2015; Wardrop et al., 2018). They could, for example, facilitate efficient delivery of public goods and services and infrastructure projects (Guiteras et al., 2018); track net migration patterns, especially in response to civilian conflict, political upheavals, and climate tragedies; and help better understand the impact of geographically-targeted economic policy interventions such as Special Economic Zones. Traditional population data sources do not meet these requirements, as censuses provide local population measurements infrequently, typically decennially. Although household surveys can yield more frequent population estimates, they are not representative at small administrative levels. The challenge of tracking population is particularly exacerbated in low- and lower-middle- income countries where population growth rates are high and net migration patterns are rapid (Figure 1).

Satellite imagery, in combination with survey data, has the potential to fill this gap by generating more frequent estimates of population that are representative in small areas. Satellites offer several advantages as data collection instruments. They systematically and universally capture large geographic areas, produce imagery at reasonably regular intervals, and are unaffected by topography, political climate, or conflict. The technology and market for both collecting and processing satellite-based imagery are developing rapidly, and with ever more satellites being launched, high-spatial resolution imagery is increasingly available at reasonable cost.

Although existing publicly available population products incorporate satellite data, this does not imply that they are up to date. Existing products typically rely on dasymetric mapping approaches to distribute population from the census to lower administrative units, either using equal weights, or weights based on satellite indicators and advanced statistical techniques. These "top-down" methods may not provide accurate or up-to-date population estimates for two reasons. First, the accuracy of distributing population relies heavily on the input population data, that is, on the census (Wardrop et al., 2018). However, the census itself may quickly become outdated due to frequent migration patterns, and rapid and uneven population growth.[1] Second, these methods often rely on Night Time Lights (NTL) or other low-spatial resolution or coarse built-up area measures from satellite and other sources to distribute census population, and hence are limited at fine spatial scales (Vogel et al., 2018).

This study seeks to overcome these limitations by proposing a "bottom-up" method which

[1] Wardrop et al. (2018) provide examples of censuses that are outdated in countries such as Lebanon and Somalia due to conflicts, and postponed or cancelled in some others such as Afghanistan, Madagascar and the Republic of Congo. It is especially critical that we are able to obtain local population estimates for conflict ridden and unstable countries not only to track settlements and refugee movements, but also to provide aid and assistance.

combines survey and satellite data to generate local estimates of population density, and by extension, population counts. Instead of relying on the census, which becomes outdated over time, the proposed method exploits the updated demographic information in surveyed areas from existing periodic household surveys, and the widespread coverage and granular information offered by satellite imagery. The method predicts population density in non-surveyed areas, and in between-census years by employing updated survey and satellite data. This technique is applied in the context of Sri Lanka using the Household Income and Expenditure Survey (HIES), a nationally-representative household survey. The satellite indicators include those derived from both low- and high-resolution satellite imagery for the entire country, and additionally, object and contextual features derived from very-high resolution imagery for a randomly selected portion of the country. These indicators are used to predict population density at the Gram Niladhari (GN) division, the lowest administrative level in Sri Lanka. The GN division is similar in size to a village in many developing country settings, and we henceforth, for ease of exposition, refer to GN divisions as "villages".

To motivate this approach, we begin by documenting the inconsistency of existing "top-down" population products at the village level, both with each other and with the census. We then address three questions that shed light on the ability of indicators derived from satellite imagery to predict population density. First, how accurately do satellite data predict census population density at the local level? Second, does including indicators derived from high and very high-resolution satellite imagery substantially improve the predictive power of the model? Third, how does the size of the training data affect prediction accuracy? We then implement the "bottom-up" approach, and finally ask how accurate are out-of-sample predictions that are derived from the HIES and satellite-imagery-based model, compared to the existing "top-down" products?

The "top-down" products considered include, WorldPop for the years 2010 and 2015, the Global Human Settlement Layer (GHSL) for 2014, High Resolution Settlement Layer (HRSL) created by Facebook for 2015, the Center for International Earth Science Information Network's (CIESIN) Gridded Population of the World (GPW) for 2010 and 2015, and LandScan for 2010. We use simple statistical measures of association to confirm that top-down estimates are poorly correlated with each other and with the census at the village level. WorldPop 2015 and Facebook are the exceptions, because they use high-resolution satellite imagery and are calibrated to the latest census data at geographically fine levels.[2] However, since even the most accurate population products use the census to redistribute population, they may quickly become outdated as the census ages, necessitating "bottom-up" methods to track changes more frequently.

[2] At Divisional Secretariat level (one level above the village), however, all estimates are highly correlated with each other and with the census, implying that accuracy at the coarser levels is easier to achieve.

We obtain satellite indicators from a variety of sources, which we categorize into four types, based on their resolution, availability, and accessibility. The publicly available indicators obtainable for the entire country are categorized into, (i) low-resolution indicators, based on imagery resolution ranging from about 0.5km per pixel to about 30m per pixel; and (ii) high-resolution imagery whose resolution ranges from about 30m per pixel to 12m per pixel. The proprietary indicators are taken from Engstrom et al. (2017), and are based on very-high resolution imagery acquired from DigitalGlobe covering 55 Divisional Secretariat divisions (one level higher than the village, henceforth referred to as sub-districts). These are divided into two categories: (iii) contextual feature indicators; (iv) object-based indicators quantifying objects such as roofs and cars.

Our results that use Poisson regressions based on variables selected by the least absolute shrinkage and selection operator (LASSO) indicate that satellite indicators accurately predict village-level census population density. Models using publicly available satellite indicators in a full national sample predict density with an out-of-sample R^2 of 0.75. Including the very-high resolution indicators and estimating the model in the 55 sub-district sample, the out-of-sample R^2 rises to 0.83. Although the public indicators perform well in predicting density in rural as compared to urban areas, the urban-sector prediction accuracy increases substantially (about 0.13 points) when object and contextual features from Engstrom et al. (2017) are added, reflecting in part the heterogeneity in the relationship between population density and built-up area in urban areas.[3] The accuracy of the predictions does not meaningfully change when the size of the training sample is reduced. Results from random forest models produce similar results to the LASSO-based regressions, and corroborate the results across different types of high-resolution resolution-types and sectors.

Finally, we obtain results from a similar Poisson model that uses estimates of population density obtained from the 2012/13 HIES instead of the census. We then compare the out-of-sample predictions from this model with the actual census population densities in villages not covered by the HIES. This yields an out-of-sample R^2 of 0.79, Spearman Rank Correlation (SRC) of 0.91, a mean Relative Error (RE) of 37%, and a median RE of 28%. Based on R^2 and mean RE, our model is more accurate than the top-down measures: GPW (both years), GHSL, LandScan, and the 2012 WorldPop, and not as accurate as Facebook and WorldPop 2015.[4] The results on these relative accuracies extend to village-level population-count predictions derived by multiplying density predictions by village area. The prediction accuracy remains similar if only indicators that are publicly available are used in the model.

[3] Rural buildings are mostly residential, but urban buildings tend to contain a heterogeneous mix of commercial and residential buildings. High resolution built-up area measures and object identifiers are better able to capture these heterogeneities in urban areas.

[4] Using median RE as the benchmark, we report comparable performances to all 2015 estimates, except Facebook and WorldPop 2015.

Two aspects of these results are particularly noteworthy. First, the survey used to calibrate the model covers only 17% of the villages in Sri Lanka. The fact that the combination of satellite imagery with this type of small survey predicts population density as accurately as the estimates that use an entire census demonstrates the cost-effectiveness of this approach.[5] In fact, simulations indicate that combining survey data with satellite data generates sub-district estimates of population density that are as precise as a survey that samples 80 percent of the villages nationwide. Second, although our estimates are not as accurate as the most accurate top-down estimates, the model gives comparable or better performance than most others. However, because surveys are collected much more frequently than censuses, bottom-up methods that combine satellite imagery with surveys have the crucial advantage of remaining up to date even as the census ages.[6]

This study fits into a rapidly growing literature on using satellite imagery to predict population counts and density in local areas. Wardrop et al. (2018) provide an overview of these studies. Top-down approaches to distribute population remain popular in the literature. For example, Stevens et al. (2015) use a dasymetric approach to redistribute population counts from the census using a Random Forest model-based weighting scheme in Cambodia, Vietnam, and Kenya. Using a combination of regression and tree-based methods, Anderson et al. (2014) predict population density for districts in Peru where no direct samples were available, using two coarse satellite-based covariates: Normalized Difference Vegetation Index (NDVI) and the daytime Land Surface Temperature (LST).[7]

Our contribution to this literature is threefold. First, relatively few studies use "bottom-up" techniques (Wardrop et al., 2018). Early studies using bottom-up style methods either use simple area measures and/or coarse satellite imagery, or mostly focus on developed countries where there are fewer data limitations (Sutton et al., 2001; Biljecki et al., 2016; Li and Weng, 2005).[8] Second, we compare prediction accuracies between imageries of various resolutions to ascertain if, and by how much, higher resolution imageries are better at predicting density compared to lower resolution imageries. Prior studies have not examined the trade-offs across

[5] It is important here to clarify that even though we do not directly rely on the census, the census remains essential to provide the sampling frame for subsequent and periodic surveys. Our method, hence, does not reduce the importance of the census.

[6] For instance, as we will show in section 4, WorldPop 2010 that utilizes the 2001 Sri Lankan census (a gap of nine years) for redistribution, performs much worse than WorldPop 2015 that uses the 2012 census (a gap of three years).

[7] We also speak to a parallel literature that uses other sources of big data such as mobile phones to map population distribution (Deville et al., 2014) or estimate population maps based on age structure (Alegana et al., 2015), and to an emerging literature that uses satellite data and machine learning techniques to predict human welfare and demographic variables, and urban market boundaries (McBride and Nichols, 2016; Athey, 2017; Engstrom et al., 2017; Jean et al., 2016; Vogel et al., 2018).

[8] Hillson et al. (2014) use bottom-up techniques but answer a different but related question of what survey size would support a particular confidence level in the estimating population using rooftop area in Bo, Sierra Leone.

different types of resolutions within the same context.

Third, we validate out-of-sample predictions derived from a household survey against the census. Such validations are not commonly performed by studies using "bottom-up" methods (Wardrop et al., 2018), except Biljecki et al. (2016) and Harvey (2002). The estimates are more accurate than previous studies that make use of two-dimensional satellite data. Our out-of-sample R^2 of 0.78 is higher than the 0.72 reported for Australia (Harvey, 2002). Our median RE at 28% is lower than those reported by Biljecki et al. (2016) in the Netherlands, which ranges from 42% to 85.4%. These higher errors could be attributed to their usage of fewer satellite indicators, linear (rather than non-linear) regressions and simple (rather than stratified) random sampling for the survey.[9]

Overall, the results demonstrate the feasibility of national statistics offices utilizing geo-spatial data in combination with frequently conducted surveys to produce accurate local population estimates in the between-census years and in areas where surveys are not conducted. Improved availability of such statistics would provide useful inputs into a wide variety of policy decisions. The rest of the paper is organized as follows: section 2 describes the data, section 3 presents the model, section 4 presents the results, and section 5 finally concludes the paper and discusses policy implications.

2 Data sources and description

2.1 Population sources

Our main measures of population density at the village level are the Census of Population and Housing, Sri Lanka, 2012, and the 2012-13 Sri Lankan Household Income and Expenditure Survey (HIES). The HIES is a detailed survey of about 25,000 households in all the districts and sub-districts of Sri Lanka, but only covers about 2,421 of the total 14,103 villages of the country. The HIES follows a two-stage stratification process: the census blocks form the Primary Sampling Units (PSUs), and the households form the Secondary Sampling Units (SSUs) or Final Sampling Units (FSUs).

The PSUs used in the HIES are census blocks, which are portions of villages.[10] Therefore, the HIES can yield direct density estimates only at the PSU-level, and not at the village-level. Unfortunately, it is not possible to model population density at the PSU-level due to the lack of a PSU-level Geographical Information System (GIS) boundary file. We therefore

[9] Interestingly, Biljecki et al. (2016) find lower errors (18.3% or 18.5%) by utilizing floor-space or volumetric 3D data and functional information about buildings. But these 3D models and data are costly to obtain periodically in low-income countries.

[10] A PSU is contained within the village. There were 62,571 PSUs versus 13,984 GNs (villages) in the 2012 census.

indirectly estimate the village population density using the HIES survey weights, under the assumption that the survey weights reflect the inverse probability of housing unit being selected for the sample. The steps used to estimate village level population density from the HIES are described in Appendix A.

2.2 Publicly available areal estimates of population density

We validate several publicly available gridded population data sets against each other and with the census, and compare them with our survey-based population density predictions. We examine the following five sources: (1) Landscan for the year 2010; (2) WorldPop for 2010 and 2015; (3) Facebook HRSL for 2015; (4) Gridded Population of the World v3 (year 2010) and v4 (year 2015), and (5) GHSL for 2015. All sources employ a top-down approach using a combination of areal interpolation techniques, including basic dasymetric approaches in conjunction with ancillary data, and statistical modeling methods to distribute census population to smaller grids. While GPW redistributes data based on an equal weighting technique, other sources use measures such as built-up area, road count and density, elevation, slope, and light intensity ("covariates") to proportionally redistribute population. Facebook, WorldPop 2015, GPW 2015, and GHSL estimates are based on the 2012 Sri Lankan census; WorldPop 2010 and GPW 2010 are based on the 2001 census; and Landscan uses mid-year population of the country of the past year calculated by the Geographic Studies Branch, US Bureau of Census (Oakridge National Laboratory, 2018).[11]

2.3 Satellite data

We use various sources of geo-spatial data in our models. The low-resolution public indicators are from three sources: (1) Night time lights from the Visible Infrared Imaging Radiometer Suite (VIIRS) at a resolution of 750 m per pixel. We use the maximum and mean intensity of two months, namely March and September, 2014; (2) Global Forest Change data based on Hansen et al. (2013), from which we use mean tree cover in 2000, and gain and loss in forest area between 2000 and 2014, at a resolution of 30 meters per pixel; (3) Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER)'s elevation and slope data at a resolution of 30 meters per pixel.

We also use publicly available built-up area measures based on higher resolution imagery. The Global Urban Footprint (GUF) (year 2012) and Global Urban Footprint plus (GUF+)

[11] Detailed information on the input population, ancillary data, and the redistribution methodology for each source is presented in the supplementary Table S1.

(year 2015) provide built-up estimates at a resolution of 12 meters per pixel.[12] Global Human Settlement Layer's (GHSL) built-up estimates are at approximately 38 meters per pixel. We also use Facebook's High Resolution Settlement Layer (HRSL) built-up area measure, available at a resolution of 30 meters per pixel.

Additionally, we use contextual and object features from Engstrom et al. (2017) which used proprietary imagery for 1,360 villages within 55 sub-districts obtained from DigitalGlobe. These images cover 3,500 km^2 of area in Sri Lanka, and are mostly obtained for the years 2011 and 2012, although some images were also captured for the year 2010. The list of 55 sub-districts is available in Appendix B. Object features extracted include the number of cars, building count and size, roof type, shadow pixels, road length and type, type of farm land (paddy or plantations).[13] Further, seven contextual features are calculated: Fourier transform, Gabor filter, Histogram of Oriented Gradients (HoG), Line Support Regions (LSR), Pantex, Normalized Difference Vegetation Index (NDVI), and Speed-Up Robustness Features (SURF).[14] More details on these objects and features can be found in Engstrom et al. (2017).

Table 1 concisely presents the key variable types, their source, time frame, resolution type, and geographical coverage. Tables 2 and 3 present the summary statistics for the population, geographic, and satellite imagery based variables for the national sample and the 55 sub-districts, respectively.

3 Modeling

3.1 Predicting population density

We use Poisson regressions to model population density at the village level. This takes the following form:

$$P_v = exp(\beta_0 + \beta_1 X_v + \beta_2 Ln\ area_v + \beta_s + \beta_d), \tag{1}$$

where P_v is the population density (persons/km^2) of village v. X_v consists of the set of satellite imagery-based indicators defined for village v. We include the natural logarithm of village area ($Ln\ area_v$), and an indicator for urban villages (ρ_S).[15] η_d represents binary indicator

[12] While GUF was made with TerraSAR-X/TanDEM-X, which are radar data sets, GUF+ adds Sentinel and Landsat which are optical remotely sensed data which substantially improves the estimation.

[13] Supplementary Figure S1 shows examples of developed area building classification, with the raw image (left) and CNN based building classification (right) in Panel A, and a sample car classification in Panel B. In a test sample, the average precision of the building classification was 71 percent, indicating a moderate share of false positives.

[14] The contextual and object features are identified using both deep learning-based Convolutional Neural Networks (CNN) and classification of spectral and textural characteristics (Engstrom et al., 2017).

[15] While the official sectoral classifications in Sri Lanka include three categories: the urban, rural, and estate sectors, we combine the second and the third to represent two categories: (1) Urban (2) Rural and estate.

variables for Sri Lanka's districts.

To prevent model over-fitting in using the entire set of satellite imagery in X_v, we employ the least absolute shrinkage and selection operator (LASSO) regularization which estimates a regression model with an added constraint that enforces parsimony. We follow a two-step procedure. First, a full set of variables is included to conduct LASSO regularization to choose variables. Second, the final Poisson model is estimated using the chosen variables. Equation 2 presents the objective function for LASSO regularization in a model using intercept β_0 and the vector of predictor coefficients τ (of length p) in step 1.

$$\min_{\beta_0,\tau} -\frac{1}{N} l(\beta_0, \tau | X, Y) + \lambda \sum_{j=1}^{p} |\tau_j|, \tag{2}$$

where l is the value of the log likelihood function of the Poisson model using parameters β_0 and τ, and λ is a non-negative regularization parameter. While setting $\lambda = 0$ yields unconstrained Poisson regression estimates, a large λ penalizes the absolute values of the coefficients, β.[16] Fivefold cross validation is applied to choose the value of λ that minimizes the root-mean squared error (RMSE) across the folds. The in-sample R^2 from this step indicates the goodness of the model fit. Since LASSO also generally shrinks the magnitude of all coefficients towards zero (Varian, 2014), we avoid biased predictors by running a simple Poisson model based on the Lasso selected variables. From this, we obtain the out-of-sample R^2, mean absolute error (MAE), and RMSE, all of which indicate out-of-sample predictive accuracy.

First, we implement the above model using the village level census population density as the dependent variable. We conduct this exercise for the entire country separately using different imagery resolution types: (1) low-resolution publicly available indicators; (2) higher resolution publicly available indicators plus the indicators mentioned in (1). Doing this enables us to compare prediction accuracy between higher versus lower resolution data in the national sample. In the 55 sub-districts where the features from Engstrom et al. (2017) are available, we implement a model using: (3) spectral and texture-features plus indicators used in (2); and (4) object-features plus the indicators used in (3). For the 55 sub-districts, we compare prediction accuracy successively from including the set of imagery-based indicators in (1)

[16] For estimations, we use the glmnet package in R that estimates the model parameters by minimizing the penalized log likelihood function. Glmnet provides an option to penalize the minimand based on the sum of the absolute values of the parameters (LASSO), or based on the sum of squares of the parameters (ridge), or a combination of both (Hastie and Qian, 2018). Glmnet's minimization function is presented in Equation 3. We set $\alpha = 1$ to use the LASSO method.

$$\min_{\beta_0,\tau} -\frac{1}{N} l(\beta_0, \tau | X, Y) + \lambda((1-\alpha) \sum_{j=1}^{p} \frac{\tau_j^2}{2} + \alpha \sum_{j=1}^{p} |\tau_j|). \tag{3}$$

through (4). We also repeat the analysis separately for urban areas and for rural and estate areas (grouped together) to check if prediction accuracy varies across sectors. Finally, focusing on the 55 sub-districts, we reduce the training sample size by half and quarter, to examine if the prediction accuracy is sensitive to the size of the training sample. As a robustness check, we also implement a flexible random forest algorithm using population density as the response variable, using the four sets of satellite indicators described above. For each model, we report out-of-bag R^2, RMSE, and MAE.[17]

Next, we focus on an overlapping sample between the sampled villages in the HIES and those in the 55 sub-districts for which proprietary features are available. These amount to 1,360 villages. In 414 villages of this sample, we model the HIES population density as a function of the all public and proprietary satellite indicators, and apply the procedure described above. Since this is a survey-based model, we use the inverse of village population for village v as weights, which is given by $\frac{1}{Village\,Population_v}$. In addition, there could be factors correlated with satellite-derived indicators that affect the probability of selection of the village into the HIES. We therefore adjust the weights based on the predicted probability of the village being sampled in HIES, based on satellite indicators available for all villages (Horvitz and Thompson, 1952; Wooldridge, 2002). This probability is obtained from the following probit model:

$$INHIES_v = \alpha + \beta lasso_v + \lambda_v, \tag{4}$$

where $INHIES_v$ is a binary indicator for whether village v is sampled into HIES, and $lasso_v$ represents all LASSO-selected variables from the Poisson model. We estimate this model using data for all 1,360 villages in the 55 sub-districts. The predicted probability of the village selection into the HIES, $\widehat{INHIES}_v$ is used for correcting the weights in the following way : $\frac{1}{VillagePopulation\,X\,\widehat{INHIES}_v}$.

Using these above two sets of weights, we estimate the model in HIES-villages. Then, we predict out-of-sample densities for the remaining 946 non-HIES villages and report out-of-sample accuracy measures by comparing them with the actual census density. To provide additional context, we compare the accuracy of this model to predictions from a similar census-based model and with density estimates from the publicly available "top-down" population products. Further, we calculate population count predictions by multiplying these density predictions with village-level area, and report their accuracy with respect to the census population.

[17] Out-of-bag R^2 is an out-of- sample statistic that refers to the average variance explained by the random forest predictions, across all villages, when averaging over the bootstrapped samples that did not contain each village.

4 Results

The sub-sections below present (a) the results comparing publicly available "top-down" population estimates with each other and with the census; (b) the predictive accuracy of satellite imagery-based indicators for census density, using LASSO-based Poisson regression models; and (c) out-of-sample density predictions based on a model using the survey (HIES), and comparing their accuracies with other estimates.

4.1 Validation of population products

The validation exercise indicates that publicly available population count estimates are accurate at the sub-district level but not necessarily at the village level. At the sub-district level, except LandScan and WorldPop 2010, the R^2 for all other products with the census are above 0.9 (panel A in Table 4). At the village level, panel B in Table 4 shows that WorldPop 2015 has the highest association with the census (R^2 of 0.988), followed by Facebooks HRSL (0.841), GHSL (0.672), GPW 2015 (0.595), GPW 2010 (.404), WorldPop 2010 (0.107) and LandScan (0.0006) in that order. These results qualitatively hold if we use correlation coefficients as a measure of association (see supplementary tables S2 and S3). Clearly, the products using the most recent census and using higher resolution imagery-based covariates are performing better than the others. For example, GPW and WorldPop 2010 estimates use the 2001 census for redistribution, and hence it is not surprising that they perform poorly in comparison to the 2012 census.[18] However, even the best performing products may themselves become inaccurate with the passage of time as their source data in the census becomes outdated.[19]

[18] Similarly, an earlier unpublished version of Facebook's population data which we worked with, only yielded a correlation coefficient of 0.5 with the 2012 Sri Lankan census at the village level because Facebook initially used the 2001 Sri Lankan census for calibration. Their estimates were later updated by using the most recent census, which now gives an R^2 of 0.841 and a correlation coefficient of 0.917.

[19] We also assess the consistency of four publicly available built-up area measures, namely, GUF, GUF+, GHSL, and Facebook, against each other, and the validate the built-up area estimates for the 55 sub-districts by Engstrom et al. (2017) against Facebook's. All estimates of built-up area at the village level are reasonably consistent with each other. At the national level, the R^2 between estimates from Facebook and each of the first three sources are 0.76 or greater (supplementary Table S4). In the 55 sub-districts, the R^2 between Facebook estimates and those from Engstrom et al. (2017) is also high at 0.794 (supplementary Table S5). These patterns also hold if correlation coefficients are employed for the analysis (see supplementary tables S6 and S7).

4.2 Accuracy of population density predictions based on census population counts

Poisson regression results using the census data for all villages, based on the variables selected from LASSO regularization indicate that satellite indicators have exceptionally strong predictive power in predicting village level population density (Table 5). At the national level, publicly available indicators explain a large amount of variation in village population density. In this case, the value added in using high-resolution indicators (out-of-sample R^2 is 0.75) as opposed to low resolution indicators (out-of-sample R^2 is 0.702) is minimal. In the 55 sub-districts, adding proprietary contextual features to public imagery-based models does not improve the prediction accuracy (out-of-sample R^2 remains at around 0.7). However, adding object classifiers to the model improves the out-of-sample R^2 by 10 points (0.83).[20]

The results differ across sectors. Using the full national sample, the publicly available satellite indicators explain more variation in population density in the rural and estate sectors (0.808) than the urban sector (0.569) (panel A in Table 6). In the sample with 55 sub-districts, the predictive power of urban population density increases tremendously when adding the proprietary contextual and object features, relative to only including publicly available indicators (0.804 compared to 0.525). Adding proprietary indicators leads only to a limited improvement in rural areas (0.869 compared to 0.838) (panel B in Table 6). This result is consistent with greater uniformity in the relationship between built-up area and population density in rural areas, where simple measures of built-up area are sufficient to predict population density. In urban areas, due to the complex nature of the relationship between population density and buildings where the latter can be both commercial and residential, advanced measures such as object classifiers and contextual measures are required for better prediction.

Random forest (RF) models provide similar results overall and across sector-types (Table 7). The out-of-bag R^2, RMSE, and MAE obtained from the RF models are comparable to the corresponding statistics obtained from the post-LASSO Poisson models. The accuracy pattern across resolution types seen in the LASSO results is also maintained in the RF results. Strikingly, the tremendous improvement in predictive accuracy from using very high resolution indicators in urban areas is corroborated by the RF results. These results collectively indicate that LASSO selection performs as well as random forest models in this setting.[21]

[20] The out-of-sample R^2 obtained in Table 5 are similar to those obtained in models that do not include log village area (supplementary Table S8), especially in models using higher-resolution indicators. The marginal effects of LASSO-selected variables from the model originally using all public and proprietary indicators are provided in supplementary Table S9. This refers to the model which uses the 55-sub-district sample and *all* imagery in Table 5.

[21] While random forest models are known to outperform LASSO models in some contexts (Mullainathan and Spiess, 2017), these results are based on linear LASSO models. The use of Poisson regression in our context may negate the predictive advantage of random forest models, which are traditionally better able to account

The size of the training data does not significantly affect the accuracy of the predictions (Table 8). Using a sample of 1,178 villages in the 55 sub-districts, the out-of-sample R^2 using all the available satellite data is 0.842, as seen earlier in Table 5. The accuracy increases marginally to 0.843 if we reduce the training sample by half, with a mild reduction to 0.832 if we further reduce the sample size by another one-half. There is only a small increase in the MAE and RMSE with the reduced sample sizes. Overall, these results indicate that changing the training sample size does not significantly change the prediction accuracy.

4.3 Survey-Based Predictions and Validation with the Census

While it is encouraging that satellite data accurately predict census population density, they do not conclusively show that survey data can be used to accurately approximate population density. We now turn to examining this by testing the accuracy of a survey-based model in the 55 sub-districts sample where we estimate a Poisson model of HIES population density in HIES-villages, and obtain out-of-sample density predictions in the non-HIES villages.

The results from a model using both public and proprietary satellite imagery-based indicators are provided in panel A, columns (1)-(3), in Table 9. The R^2, Spearman Rank Correlation (SRC), and Mean Absolute Error (MAE) between the density predictions and the census density are 0.79, 0.91, and 664, respectively. This model uses the inverse of village population as weights. If the weights are corrected based on probability of the village being selected into the HIES, we obtain similar results: 0.78, 0.92, and 665 for the same measures. Considering the R^2 and SRC, our model performs equally well compared to the census based model, and better than all "top-down" estimates except Facebook and Worldpop 2015 (panel C).[22] In panel B, we present the same results based on a model using only publicly available satellite imagery. Interestingly, similar out-of-sample prediction accuracies are obtained using only public imagery (0.75, 0.92, and 663, and 0.77, 0.92, and 657 for the three measures using the two types of weights).

We multiplied these density predictions with village-level area to obtain population count predictions, which are reported in columns (4)-(6). The relative performance of the population count predictions, in comparison to the top-down estimates, is similar to those of the population density predictions. The R^2 and SRC are generally lower for population counts compared to density predictions, but the MAEs are also lower.

for non-linear relationships. Further, studies predicting population counts and density have specifically shown mixed results on comparative performances between random forest and LASSO-based predictions (Anderson et al., 2014).

[22] The out-of-sample R^2, SRC, MAE, in that order, for the "top-down" measures are the following : WorldPop 2015 (0.99, 1.00, and 312) and Facebook's HRSL (0.91, 0.97, and 417), GHSL (0.84, 0.88, and 639), GPW 2015 (0.68, 0.90, and 1029), GPW 2010 (0.29, 0.28, and 1550), WorldPop 2010 (0.31, 0.81, and 1206), and LandScan (0.04, 0.10, and 2562).

The mean and the median relative errors (which are same for population density and count predictions) are reported in columns (7) and (8), respectively.[23] The mean and median relative errors of the model are 37% and 28%, and 35% and 29%, without and with the additional weight correction. Using only public imagery, the mean and median REs are 34% and 28%, and 34% and 28% without and with weights correction, respectively. These mean REs are lower than most top-down estimates, except Facebook and WorldPop 2015. The median REs are higher than only Facebook and Worldpop 2015, and comparable to GHSL and GPW 2015 and much lower compared to GPW, Worldpop, and Landscan (all in 2010).[24]

The coefficient of variation (CV) of predicted density estimates for each sub-district based on the HIES model in combination with satellite imagery is low, at about 11%. This is four times lower than the CV of the survey estimates themselves, which stands at 45.2%. The increase in precision due to the incorporation of satellite imagery is tantamount to increasing the share of sampled villages in the population from 17%, which is currently the case, to a remarkable 80%.[25] This is striking but not surprising because the model predicts with an out-of-sample R^2 of about 0.80 and the imagery covers all villages.

5 Discussion and Conclusion

Existing methods of estimating local population use a "top-down" dasymetric mapping approach to distribute census data based on a set of covariates derived from satellite imagery. These are only as accurate as the source data—the census—and hence their accuracy declines as the census itself ages. We propose a "bottom-up" technique by pairing survey data with satellite imagery using Poisson regressions employing variables selected based on LASSO regularization. This model can yield population density predictions in areas where the survey was not conducted, and can be repeated frequently using periodic survey and satellite data to frequently track population changes. We apply the method in the context of Sri Lanka to predict population density at the lowest administrative level, the Gram Niladhari (GN) division (village-level). Our model uses the Sri Lankan Household Income and Expenditure Survey (HIES), and predicts density out-of-sample in the non-HIES villages.

Two aspects of our results stand out. First, our performance is better than the only

[23] The relative error with respect to the census is the same for population density and population count because the only difference between the two is the multiplicative factor, village-area.

[24] We conducted a similar exercise using population count as the dependent variable in the model, from which we directly obtained population predictions. The accuracies of these predictions are reported in Table S10. Examining the R^2, SRC, and Mean RE, it is evident that the population predictions derived indirectly using a density-based model by multiplying density predictions with village-area (as in Table 9) performs much better than obtaining population count predictions from a population-based model.

[25] The CV of 45.2% was obtained by manually calculating the mean CVs from a synthetic 80% sample of the census.

two other studies that report out-of-sample validation of "bottom-up" population estimations using similar satellite data (Harvey, 2002; Biljecki et al., 2016). Second, our predictions outperform many "top-down" estimates with the exception of Facebook and WorldPop 2015. Although ours is not the most accurate, it crucially does not *directly* rely on the census data.

The main reason why this method is useful is because surveys are more frequent and less expensive than censuses, and satellite data can be acquired routinely and with complete spatial coverage, and hence can potentially be more efficiently utilized to track population changes in small areas. Frequent up-to-date small area population estimations are important inputs for governments in applications such as tracking disaster management, delivering policy programs, understanding migration patterns, and many more. The Sri Lankan Department of Census and Statistics conducts the HIES every three years; the most recent was fielded in 2016, and the one before that was in 2012-13, and the next one is planned for 2019. Furthermore, Sri Lanka runs a continual Labor Force Survey that could also be used for this purpose. Therefore, the results support the case for the Department of Census and Statistics to keep careful track of the number of buildings identified in the relisting phase, and use that information to generate revised local population estimates.

References

Alegana, V. A., Atkinson, P. M., Pezzulo, C., Sorichetta, A., Weiss, D., Bird, T., Erbach-Schoenberg, E., and Tatem, A. J. (2015). Fine resolution mapping of population age-structures for health and development applications. *Journal of The Royal Society Interface*, 12(105):20150073.

Anderson, W., Guikema, S., Zaitchik, B., and Pan, W. (2014). Methods for Estimating Population Density in Data-Limited Areas: Evaluating Regression and Tree-Based Models in Peru. *PloS one*, 9(7):e100037.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485.

Biljecki, F., Ohori, K. A., Ledoux, H., Peters, R., and Stoter, J. (2016). Population estimation using a 3d city model: A multi-scale country-wide study in the netherlands. *PloS one*, 11(6):e0156808.

Centre for International Earth Science Information Network (2018). High resolution settlement layer. Retrieved from https://www.ciesin.columbia.edu/data/hrsl/ on August 4th 2018.

Department of Census and Statistics (2015). Household income and expenditure survey 2012/13. final report. Department of Census and Statistics, Ministry of Policy Planning Economic Affairs, Child Youth and Cultural Affairs, Sri Lanka.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893.

Engstrom, R., Hersh, J., and Newhouse, D. (2017). Poverty from space: Using high-resolution satellite imagery for estimating economic well-being. World Bank Working Paper No. 8284.

European Commission (2018). Global Human Settlement (GHSL). Retrieved from https://ghsl.jrc.ec.europa.eu/data.php on November 19th 2018.

Guiteras, R., Levinsohn, J., and Mobarak, M. (2018). Demand estimation with strategic complementarities: Sanitation in bangladesh.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S., Tyukavina, A., Thau, D., Stehman, S., Goetz, S., Loveland, T., et al. (2013). High-resolution global maps of 21st-century forest cover change. *science*, 342(6160):850–853.

Harvey, J. (2002). Estimating census district populations from satellite imagery: some approaches and limitations. *International Journal of Remote Sensing*, 23(10):2071–2095.

Hastie, T. and Qian, J. (2018). Glmnet vignette. Retrieved from https://web.stanford.edu/ hastie/glmnet/glmnet$_a$*lpha.html*poi on November 19th 2018.

Hillson, R., Alejandre, J. D., Jacobsen, K. H., Ansumana, R., Bockarie, A. S., Bangura, U., Lamin, J. M., Malanoski, A. P., and Stenger, D. A. (2014). Methods for determining the uncertainty of population estimates derived from satellite imagery and limited survey data: a case study of bo city, sierra leone. *PloS one*, 9(11):e112241.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Li, G. and Weng, Q. (2005). Using landsat etm+ imagery to measure population density in indianapolis, indiana, usa. *Photogrammetric Engineering & Remote Sensing*, 71(8):947–958.

McBride, L. and Nichols, A. (2016). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Oakridge National Laboratory (2018). Documentation page of the landscan. Retrieved from https://landscan.ornl.gov/documentation on August 3rd 2018.

Socioeconomic Data and Applications Center (2018). Gridded Population of the World (GPW). Retrieved from http://sedac.ciesin.columbia.edu/data/collection/gpw-v4/methods/method1 on November 19th 2018.

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

Sutton, P., Roberts, D., Elvidge, C., and Baugh, K. (2001). Census from heaven: An estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing*, 22(16):3061–3076.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

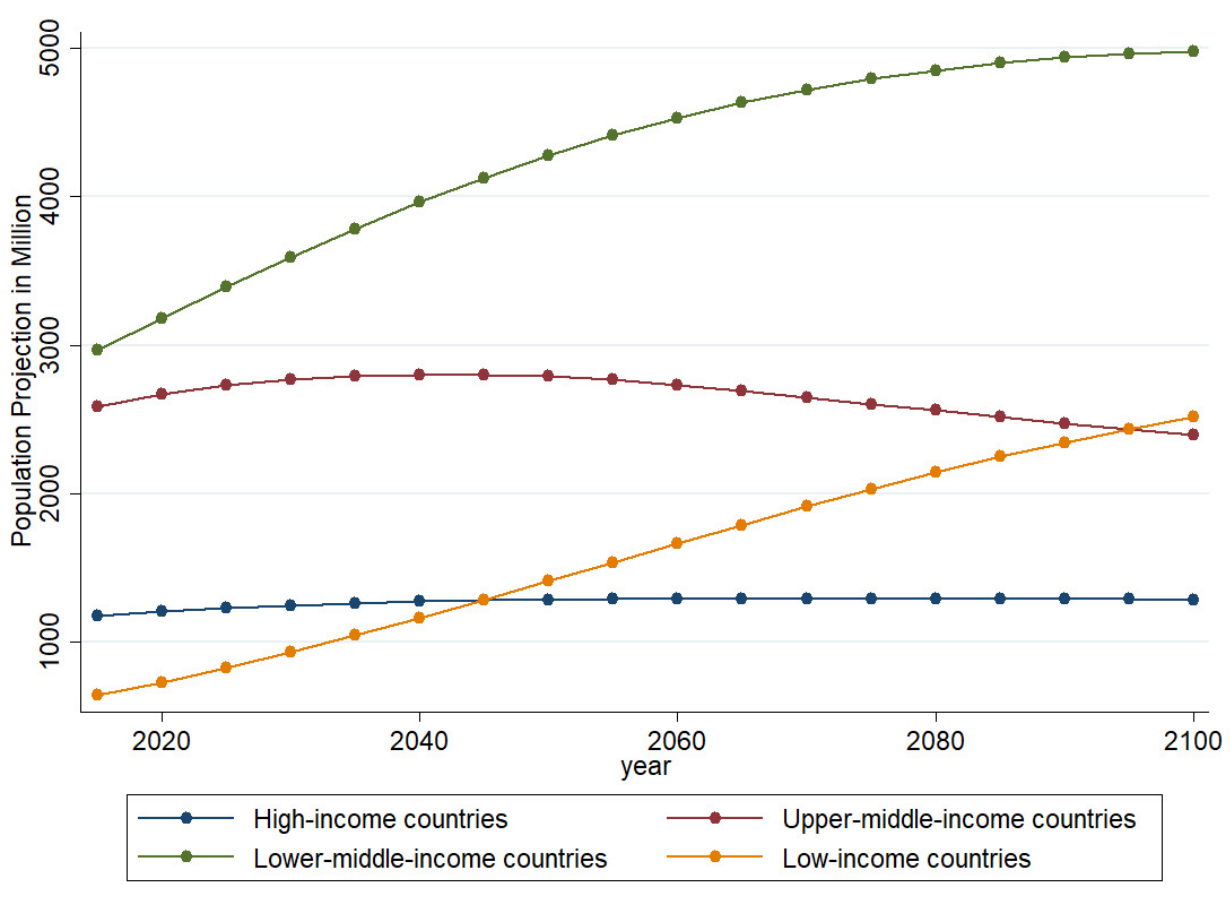
Vogel, K., Goldblatt, R., Hanson, G., and Khandelwal, A. (2018). Detecting urban markets with satellite imagery. International Growth Centre Working Paper C-89448-INC-1.

Wardrop, N., Jochem, W., Bird, T., Chamberlain, H., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., and Tatem, A. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences*, 115(14):3529–3537.

Wooldridge, J. M. (2002). Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1(2):117–139.

WorldPop (2018). World pop website and methods page. Retrieved from http://www.worldpop.org.uk/data/methods/ on August 4th 2018.

Figure 1: Total Population Projections, by Country Group



Source: World Population Prospects 2017, United Nations

Table 1: Satellite Indicators by Resolution and Availability

Indicator	Source	Time-frame	Coverage	Resolution	Availability
A. Population					
Population	Census		National		public
Population	HIES		1421 villages		public
Population	Facebook		National		public
Population	WorldPop		National		public
Population	GPW		National		public
Population	LandScan		National		public
Population	GHSL		National		public
B. Other sources					
Night time lights	VIIRS	2014	National	Low	public
Elevation	ASTER		National	Low	public
Slope	ASTER		National	Low	public
Tree cover	Hansen et al. (2013)	2000	National	Low	public
Tree cover gain and loss	Hansen et al. (2013)	2000 to 2014	National	Low	public
Built-up area	GUF, GUF+	2012, 2015	National	High	public
Built-up area	GHSL	2014	National	High	public
Built-up area	Facebook	2015	National	High	public
Built-up area	Engstrom et al. (2017)	2011, 2012	55 sub-dist.	High	proprietary
Cars	Engstrom et al. (2017)	2011, 2012	55 sub-dist.	High	proprietary
Shadows	Engstrom et al. (2017)	2011, 2012	55 sub-dist.	High	proprietary
Roof type	Engstrom et al. (2017)	2011, 2012	55 sub-dist.	High	proprietary
Road type	Engstrom et al. (2017)	2011, 2012	55 sub-dist.	High	proprietary
Agricultural land type	Engstrom et al. (2017)	2011, 2012	55 sub-dist.	High	proprietary
Paddy land type	Engstrom et al. (2017)	2011, 2012	55 sub-dist.	High	proprietary
NDVI	Engstrom et al. (2017)	2011, 2012	55 sub-dist.	High	proprietary
Other Contextual Indicators	Engstrom et al. (2017)	2011, 2012	55 sub-dist.	High	proprietary

Note: HIES=the Household Income and Expenditure Survey; GPW=Gridded Population of the World; GHSL=Global Human Settlement Layer; GUF= Global Urban Footprint; VIIRS= Visible Infrared Imaging Radiometer Suite; ASTER=Advanced Spaceborne Thermal Emission and Reflection Radiometer; NDVI = Normalized Difference Vegetation Index. Engstrom et al. (2017) use proprietary imagery from DigitalGlobe to derive features and textures for 55 sub-district divisions. "Other Contextual indicators" include SURF, Pantex, and Histogram of Oriented Gradients. The time frame refers to the year(s) during which the satellite images were obtained.

Table 2: Summary Statistics: Villages in the National Sample

Indicator	Mean	Std. Dev.	Min	Max
Geographic Descriptors				
log village area	1.25	0.81	0.04	6.35
Share of urban villages	0.092	0.289	0.00	1.00
Population summary				
Village level				
Village Census Population Density(per km^2)	1,400	2,703	0	50,126
Village HIES Population Density(per km^2; N=2,348)	1,996	3,394	0.607	49,209
Village population- Census	1,455	1,244	12	28,003
Village population estimate - Facebook	1,477	1,334	0	30,046
Village population estimate - WorldPop 2010	1,451	1,246	0	28,003
Village population estimate - WorldPop 2015	1,263	1,081	14	26,870
Village Population estimate - GPW 2010	1,387	1,652	0	38,355
Village Population estimate - GPW 2015	1,650	1,689	18	37,381
Village population estimate - LandScan	1,519	2,873	0	61,275
Village population estimate - GHSL	1,468	1,475	0	38,775
Sub-district level (N=332)				
Sub-district population- Census	61,322	48,605	0	323,223
Sub-district population estimate - Facebook	62,223	49,471	3,217	318,643
Sub-district population estimate - WorldPop 2010	61,252	42,621	4,742	261,693
Sub-district population estimate - WorldPop 2015	53,119	42,095	281	277,480
Sub-district Population estimate - GPW 2010	58,442	45,983	3,505	330,155
Sub-district Population estimate - GPW 2015	61,062	48,140	3,291	318,456
Sub-district population estimate - LandScan	64,211	44,390	0	268,650
Sub-district population estimate - GHSL	61,772	48,865	1,661	310,012
Key Satellite Imagery Based indicators				
Low Resolution Public Indicators				
Night-time lights March 2014 - Mean	0.98	2.25	0.000	72.57
Night-time lights March 2014 - Maximum	1.45	4.13	0.000	274.93
Night-time lights September 2014 - Mean	0.95	2.04	0.000	56.61
Night-time lights September 2014 - Maximum	1.39	3.28	0.0	170.62
Mean Elevation	197.13	313.35	3.4	2,214.55
Mean Slope	8.68	5.15	1.2	28.79
Mean Tree Cover	46.93	26.91	0.0	97.26
Gain in Tree Cover	0.02	0.02	0.000	0.25
Loss in Tree Cover	0.005	0.012	0.000	0.330
High Resolution Public Indicators				
GUF Built-up Area	10.60	20.37	0.0	100.00
GUF+ Built-up Area	14.21	25.83	0.0	101.05
GHSL Built-up Area	15.41	24.88	0.0	100.00
Facebook Built-up Area (N= 13,437)	15.65	19.16	0.0	99.55
Observations (N)		13,970		

Note: GPW=Gridded Population of the World; GHSL=Global Human Settlement Layer; HIES=Household Income and Expenditure Survey; GUF= Global Urban Footprint; NDVI = Normalized Difference Vegetation Index.

Table 3: Summary statistics: Villages in the 55 Sub-districts Sample

	Mean	Std. Dev.	Min	Max
Geographic Descriptors				
Log village area	0.97	0.60	0.09	4.27
Share of urban villages	0.28	0.45	0.00	1.00
Population Summary				
Village level				
Village Census Population Density	2621	3507	22	43,984
Village HIES Population Density(per km^2; N=414)	2,821	3,576	10.22	38,609
Village population- Census	2341	2183	212	28,003
Village population estimate - Facebook	2350	2176	82	30,046
Village population estimate - WorldPop 2010	2014	1928	0	21,326
Village population estimate - WorldPop 2015 (N=1,357)	2047	1919	215	26,870
Village Population estimate - GPW 2010	2192	2886	0	36,832
Village Population estimate - GPW 2015 (N=1,145)	2747	2813	146	34,058
Village Population estimate - LandScan	1359	2655	0	40,639
Village Population estimate - GHSL	1359	2655	0	40,639
Sub-district level (N=55)				
Sub-district population- Census	57,887	54,117	1,709	2,40,114
Sub-district population estimate - Facebook	58,100	55,066	1,714	2,51,530
Sub-district population estimate - WorldPop 2010	49,812	46,543	1,222	1,96,328
Sub-district population estimate - WorldPop 2015	50,494	47,191	1,436	2,07,748
Sub-district Population estimate - GPW 2010	54,212	52,689	2,025	2,28,419
Sub-district Population estimate - GPW 2015	57,181	54,766	1,655	2,41,757
Sub-district population estimate - LandScan	33,606	37,591	0.00	1,53,304
Sub-district population estimate - GHSL	57,799	55,175	1,708	2,46,673
Low Resolution Public Indicators				
Night-time lights March 2014 - Mean	1.39	3.27	0.00	57.21
Night-time lights March 2014 - Maximum	2.10	8.54	0.00	274.93
Night-time lights September 2014 - Mean	1.26	2.50	0.00	28.96
Night-time lights September 2014 - Maximum	1.76	3.88	0.00	92.13
Mean Elevation	209.96	428.52	5.56	2214.55
Mean Slope	7.63	5.14	1.15	23.89
Mean Tree Cover	44.22	26.35	0.00	92.24
Gain in Tree Cover	0.02	0.03	0.00	0.22
Loss in Tree Cover	0.01	0.02	0.00	0.33
High Resolution Public Indicators				
GUF Built-up Area	21.81	26.92	0.00	97.11
GUF+ Built-up Area	31.68	36.52	0.00	100.11
GHSL Built-up Area	33.59	35.68	0.00	100.00

Continued on next page

Table 3 – *Continued from previous page*

	Mean	Std. Dev.	Min	Max
Facebook Built-up Area (N=1,275)	30.16	28.18	0.67	97.43
Very High resolution Proprietary indicators				
Road variables				
% of roads that are minor paved (4 m width)	6.78	15.50	0.00	100
% of roads that are main paved (5 m width)	11.32	15.07	0.00	100
% of roads that are paved city (4 m width)	11.74	17.64	0.00	92.76
Building Density and Vegetation				
% shadow pixels covering valid area (N=1,353)	5.80	5.47	0.27	38.17
NDVI, mean scale 32	0.52	0.17	0.00	0.95
Total built-up area	69,757	68,699	0	7,08,867
Roof type				
Fraction of total roofs that are clay	35.80	20.67	0.00	100
Fraction of total roofs that are aluminum	14.26	7.27	0.00	71.92
Fraction of total roofs are asbestos	7.79	11.66	0.00	71.20
Cars				
log number of cars (N=1,252)	3.42	1.02	0.88	8.30
Agricultural Land				
% of Village agriculture that is paddy	44.46	37.65	0.00	100.00
% of Village agriculture that is plantation	55.14	37.58	0.00	100.00
Textural and spectral characteristics				
Pantex (human settlements) mean, scale 8m	0.56	0.50	0.00	3.95
Pantex (human settlements) mean, scale 32m	0.66	0.59	0.00	4.68
Gabor filter (scale 64m, features 6), mean	0.67	0.32	0.01	2.03
Gabor filter (scale 64m, features 14), mean	0.68	0.31	0.02	1.95
Histogram of Oriented Gradients (scale 16m), mean	37.91	9.66	0.00	146.18
Observations (N)		1,360		

Note: GPW=Gridded Population of the World; GHSL=Global Human Settlement Layer. HIES=Household Income and Expenditure Survey; GUF= Global Urban Footprint; NDVI = Normalized Difference Vegetation Index. The NDVI scale refers to the size of the window used to calculate average NDVI. Summary statistics of LASSO-selected indicators across all models are reported.

Table 4: Validation of Population Estimates

A. Sub-district Level

R^2	Census	Facebook	WorldPop 2010	WorldPop 2015	GPW 2010	GPW 2015	GHSL	LandScan
Census	1							
Facebook	.996	1						
WorldPop 2010	.753	.758	1					
WorldPop 2015	.999	.995	.757	1				
GPW 2010	.985	.982	.748	.986	1			
GPW 2015	.986	.989	.770	.988	.990	1		
GHSL	.994	.996	.766	.995	.982	.990	1	
LandScan	.015	.013	.026	.015	.022	.017	.014	1
Average discrepancy relative to census								
Persons	0	2468	15107	10360	4948	3663	2778	46936
Percent	0%	4.02%	21.03%	13.38%	6.72%	5.37%	4.79%	69.58%

B. Village Level

R^2	Census	Facebook	WorldPop 2010	WorldPop 2015	GPW 2010	GPW 2015	GHSL	LandScan
Census	1							
Facebook	.841	1						
WorldPop 2010	.107	.093	1					
WorldPop 2015	.988	.833	.107	1				
GPW 2010	.404	.466	.049	.420	1			
GPW 2015	.595	.634	081	.604	.595	1		
GHSL	.672	.801	.076	.671	.408	.527	1	
LandScan	.0006	.0009	.0008	.00007	.00007	.0009	.0007	1
Average discrepancy relative to census								
Persons	0	292	733	197	669	553	515	1684
Percent	0%	22.62%	72.12%	13.5%	51.02%	45.77%	42.42%	197.37%

Note: GPW=Gridded Population of the World; GHSL=Global Human Settlement Layer. Discrepancy is the absolute value of the difference between population estimates and the census at the village level. Percent relative to the census is (absolute-discrepancy/census-population) X 100.

Table 5: Accuracy of Population Density Predictions based on LASSO Selected Poisson Regression of Population Density

Imagery Type	In-sample R^2	Out-of-sample R^2	Number of variables		No. of Villages
			Candidate	Selected	
A. National Sample					
1. No satellite imagery	0.806	0.650	35	15	13,970
2. Low-resolution, public	0.843	0.702	35	18	13,970
3. All public	0.888	0.750	39	13	13,437
B. 55 sub-district sample					
1. No satellite imagery	0.731	0.589	35	5	1,360
2. Low-resolution, public	0.804	0.677	35	7	1,360
3. All public	0.858	0.710	39	9	1,275
4. All public + propriety texture	0.874	0.755	46	11	1,275
5. All	0.918	0.830	67	23	1,178

Note: The results are based on a Poisson regression model whose dependent variable is the census village population density, and the independent variables are selected based on LASSO regularization among the following: Relevant satellite imagery sources as described in panel B in Table 1, district fixed effects, a binary indicator for urban villages, and log-village area. The models with "No satellite imagery" use only the district fixed effects, an indicator for urban villages, and log village-area as independent variables. "Low-resolution public" indicators refer to the low-resolution and public indicators as defined in Table 1. "All public" indicators refers to all the public imagery based indicators, including the high-resolution imagery based indicators described in Table 1. "Proprietary texture" indicators include very high-resolution texture features based on Digital Globe from Engstrom et al. (2017) as defined in Table 1. "All" imagery refers to all public and proprietary texture indicators, and object identifiers, such as cars, roofs etc., as defined as in Table 1. The out-of sample R^2 is obtained from stata's crossfold command using five-fold cross-validation.

Table 6: Accuracy of Population Density Predictions, by Sector and Nature of Indicators

Imagery Type	In-sample R^2	Out-of-sample R^2	Number of variables		No. of villages
			Candidate	Selected	
		A. National Sample			
A. Rural and estate					
1. No satellite imagery	0.757	0.646	35	23	12,686
2. Low resolution, public	0.806	0.726	35	25	12,686
3. All Public	0.877	0.808	39	19	12,389
B. Urban					
1. No satellite imagery	0.595	0.493	35	3	1,284
2. Low resolution, public	0.654	0.540	35	7	1,284
3. All Public	0.690	0.569	39	6	1,048
		A. 55 sub-districts			
A. Rural and estate					
1. No satellite imagery	0.752	0.688	35	8	978
2. Low resolution, public	0.815	0.754	35	8	978
3. All Public	0.898	0.838	39	8	935
4. All Public + Propriety texture	0.919	0.882	46	24	935
5. All	0.942	0.869	67	41	881
B. Urban					
1. No satellite imagery	0.487	0.413	35	3	382
2. Low resolution, public	0.637	0.511	35	5	382
3. All Public	0.665	0.525	39	7	340
3. All Public + Propriety texture	0.740	0.668	46	8	340
5. All	0.860	0.804	67	26	336

Note: The results are based on a Poisson regression model whose dependent variable is census village population density, and the independent variables are selected based on LASSO regularization among the following: Relevant satellite imagery sources as described in panel B in Table 1, district fixed effects, a binary indicator for urban villages, and log village area. The models with "No satellite imagery" use only the district fixed effects, an indicator for urban villages, and log village-area as independent variables. "Low-resolution public" indicators refer to the low-resolution and public indicators as defined in Table 1. "All public" indicators refers to all the public imagery based indicators, including the high-resolution imagery based indicators described in Table 1. "Proprietary texture" indicators include very high-resolution texture features based on Digital Globe from Engstrom et al. (2017) as defined in Table 1. "All" imagery refers to all public and proprietary texture indicators, and object identifiers, such as cars, roofs etc., as defined as in Table 1. The out-of sample R^2 is obtained from stata's crossfold command using five-fold cross-validation.

Table 7: Random Forest model Results in comparison with LASSO-selected Poisson Regression Results

	Random Forest			LASSO		
Imagery Type	Out-of-bag R^2	RMSE	MAE	Out-of-sample R^2	RMSE	MAE
All sectors						
A. National sample						
1. Low resolution, public	0.767	1282	477	0.702	1463	563
2. All Public	0.783	1237	390	0.750	1323	459
B. Subsample of 55 sub-districts						
1. Low resolution, public	0.677	2088	948	0.677	1948	928
2. All Public	0.707	1984	784	0.710	1927	818
3. All Public + Propriety texture	0.719	1942	779	0.755	1845	808
4. All	0.792	1673	684	0.830	1506	698
Rural and Estate						
A. National sample						
1. Low resolution, public	0.776	696	329	0.726	788	375
2. All Public	0.832	601	246	0.808	653	289
B. Subsample of 55 sub-districts						
1. Low resolution, public	0.807	968	510	0.754	990	550
2. All Public	0.847	855	383	0.838	848	424
3. All Public + Propriety texture	0.849	850	380	0.882	788	415
4. All	0.872	781	352	0.869	747	355
Urban						
A. National sample						
1. Low resolution, public	0.598	3966	2285	0.540	4105	2344
2. All Public	0.612	3890	2079	0.569	3999	2343
B. Subsample of 55 sub-districts						
1. Low resolution, public	0.424	3680	2084	0.511	3277	1917
2. All Public	0.461	3559	1888	0.525	3357	1909
3. All Public + Propriety texture	0.491	3457	1850	0.668	2897	1676
4. All	0.626	2984	1620	0.804	2226	1446

Note: Random Forest models are implemented using the randomForest() package in R with 500 trees, and using the census village population density as the dependent variable. The following independent variables were used in all models: Different types of satellite imagery sources as described in panel B in Table 1, district fixed effects, a binary indicator for urban villages, and log village-area. "Low-resolution public" indicators refer to the low-resolution and public indicators as defined in Table 1. "All public" indicators refers to all the public imagery based indicators, including the high-resolution imagery based indicators described in Table 1. "Proprietary texture" indicators include very high-resolution texture features based on Digital Globe from Engstrom et al. (2017) as defined in Table 1. "All" imagery refers to all public and proprietary texture indicators, and object identifiers, such as cars, roofs etc., as defined as in Table 1. The out-of-sample R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) for the LASSO based Poisson regression are from tables 8-10.

Table 8: Accuracy of Population Density Predictions, by Training Sample Size

		In-sample	Out-of-sample		
Sample Type	Number of Villages	Pseudo R^2	Pseudo R^2	MAE	RMSE
Medium sample	1178	0.918	0.842	1490	712
Small sample (50% of medium)	593	0.908	0.843	1447	698
Very small sample (50% of small)	292	0.920	0.832	1691	738

Note: The results are based on a Poisson regression model whose dependent variable is census village population density, and the independent variables are selected based on LASSO regularization among the following: All indicators from public and proprietary imagery sources as described in panel B in Table 1, sectoral and district fixed effects, and log village-area. The out-of sample R^2 is obtained from stata's crossfold command using five-fold cross-validation. MAE refers to mean absolute error and RMSE refers to Root Mean Squared Error.

Table 9: Out-of-sample Accuracy of HIES-based Village Population Density Estimates using Public and Proprietary satellite data, 55 Sub-Districts

	Population Density			Population Count				
	R^2	SRC	Mean AE	R^2	SRC	Mean AE	Mean RE	Median RE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A.Model estimates, Public & Proprietary satellite imagery								
HIES-based	0.79	0.91	664	0.60	0.67	607	37%	28%
HIES-based (balance correction)	0.78	0.92	665	0.58	0.67	607	35%	29%
Census-based	0.81	0.91	679	0.60	0.65	644	50%	28%
B.Model estimates, public satellite imagery								
HIES-based	0.75	0.92	663	0.56	0.69	595	34%	28%
HIES-based (balance correction)	0.77	0.92	657	0.55	0.69	606	34%	28%
Census-based	0.76	0.91	750	0.56	0.63	739	58%	36%
C. Existing areal interpolation estimates								
WorldPop 2015	0.99	1.00	312	0.99	0.99	237	13%	13%
Facebook HSRL 2015	0.91	0.97	417	0.90	0.93	309	20%	14%
GHSL Grid 2014	0.84	0.88	639	0.74	0.79	556	39%	26%
GPW 2015	0.68	0.90	1029	0.50	0.78	740	44%	27%
GPW 2010	0.29	0.28	1550	0.39	0.48	911	52%	35%
WorldPop 2010	0.31	0.81	1206	0.06	0.44	987	64%	34%
LandScan 2010	0.04	0.10	2562	0.01	-0.24	2012	135%	92%

Note: HIES=the Household Income and Expenditure Survey; HRSL= High Resolution Settlement Layer; GHSL=Global Human Settlement Layer; GPW=Gridded Population of the World; SRC=Spearman Rank Correlation; AE= Absolute error; RE=Relative Error (with respect to the census). We retain only the villages in the 55 sub-districts for which the proprietary data are available. Models are estimated on subsample of 414 villages covered in 2012-2013 HIES survey. The satellite indicators used as covariates include all publicly available and all proprietary indicators. Out-of-sample prediction accuracy (with the census) are conducted on the 946 villages not covered in the HIES. The weights for the HIES-based model are $\frac{1}{Village\,Population_v}$. Balance correction refers to adjustment of weights based on the predicted probability of the village being sampled in HIES ($\widehat{INHIES_v}$) based on a probit model with village sampled in HIES as the dependent variable and all LASSO-selected variables as independent variables. The balanced corrected weights are $\frac{1}{Village Population_v\, X\widehat{INHIES_v}}$. Population count predictions were obtained by multiplying density predictions by village-level area. The accuracy parameters for population predictions are calculated based on comparisons with actual census population and reported. The mean and median RE are mathematically the same for population density and population count predictions.

Appendix

A Estimating GN (village)-population from the HIES

To estimate village-level population density from the Sri Lankan Household Income and Expenditure Survey (HIES), we use a special version of the HIES obtained from the Department of Census and Statistics (DCS) that contains village identifiers. We infer the village population from sample weights using the following identity:

$$W_{h,v} = \frac{1}{prob(h|PSU)} \frac{1}{prob(PSU)}, \tag{1}$$

where $W_{h,v}$ represents the sample weight assigned to household h in a particular village v. Equation 1 reflects the standard definition of the sample weight which equals the inverse probability of a household being selected. Because of the two-stage sampling design employed by the DCS, the probability of the household being surveyed is equal to the product of the probability of the Primary Sampling Unit (PSU) being selected ($prob(PSU)$) and the probability that the household is selected conditional on the PSU being selected ($prob(h|PSU)$). The probability that the household is selected conditional on its PSU being selected is:

$$prob(h|PSU) = \frac{HU_{sample}}{HU_{psu}} \tag{2}$$

HU_{sample} indicates the number of household units in the sample, and HU_{psu} indicates the number of household units in the PSU. Equation 2 indicates that each housing unit within a PSU has an equal probability of selection. The numerator can easily be calculated from the HIES sample, while the denominator needs to be backed out from the sample weights. An important note is that the denominator is updated each time a new survey is conducted by survey teams that lists *all* households in sampled PSUs.

The second term of Equation 1 is the inverse probability that the PSU is selected. The

probability that the PSU is selected is equal to the share of census housing units in that PSU (Department of Census and Statistics, 2015). To estimate population at the village level, we utilize the following application of Bayes rule:

$$Prob(PSU) = Prob(\frac{PSU}{v}) * Prob(v) \tag{3}$$

This decomposition is a mathematical identity rather than a description of the actual sample design. It is useful because the first term in Equation 3 can be rewritten as follows:

$$Prob(\frac{PSU}{v}) = \frac{HU_{psu}}{HU_v} \tag{4}$$

This indicates that, if hypothetically the village of the selected PSU was known, the probability that the PSU was selected for the sample is the share of that village's housing units contained in that PSU. This identity holds only for villages that exactly contain one PSU, which applies to 97 percent of the villages in the HIES sample.[26] Substituting Equation 2, Equation 3, and Equation 4 in Equation 1 gives:

$$W_{h,v} = \frac{HU_v}{HU_{sample}} * \frac{1}{prob(v)} \tag{5}$$

$$HU_v = W_{h,v} * HU_{sample,v} * Prob(v) \tag{6}$$

The first term on the right-hand side of Equation 6 is the sample weight for household h in a particular village. The second is the number of households in that village that were sampled in the survey. The final term is the probability that the village is in the sample, which needs to be estimated using publicly available indicators on the population of each village in the 2011 census.[27] We merge census population with the HIES data at the village level and estimate the probability of village selection as the following:

[26] 93.5% of the PSUs in the sample are located in single PSU-villages.
[27] The village population estimates are available at LankaStatMap (http://www.map.statistics.gov.lk/)

$$Prob(v) = 1 - \prod_{i \neq v} \frac{HU_v}{HU_{country} - \sum_{j=1}^{i} HU_j} \tag{7}$$

The probability that a village was selected for the sample is equal to one minus the probability that the village was not selected. The probability that a village was not selected is equal to the product, across all other villages in the sample, of that village not being selected each time a PSU is drawn for the sample. Since the sample is without replacement, in each draw the probability that a village is not selected is equal to one minus the share of remaining non-sampled housing units in each round. We simulate this probability using the actual other villages selected for the sample. This can be calculated for each village based on publicly available village population data.

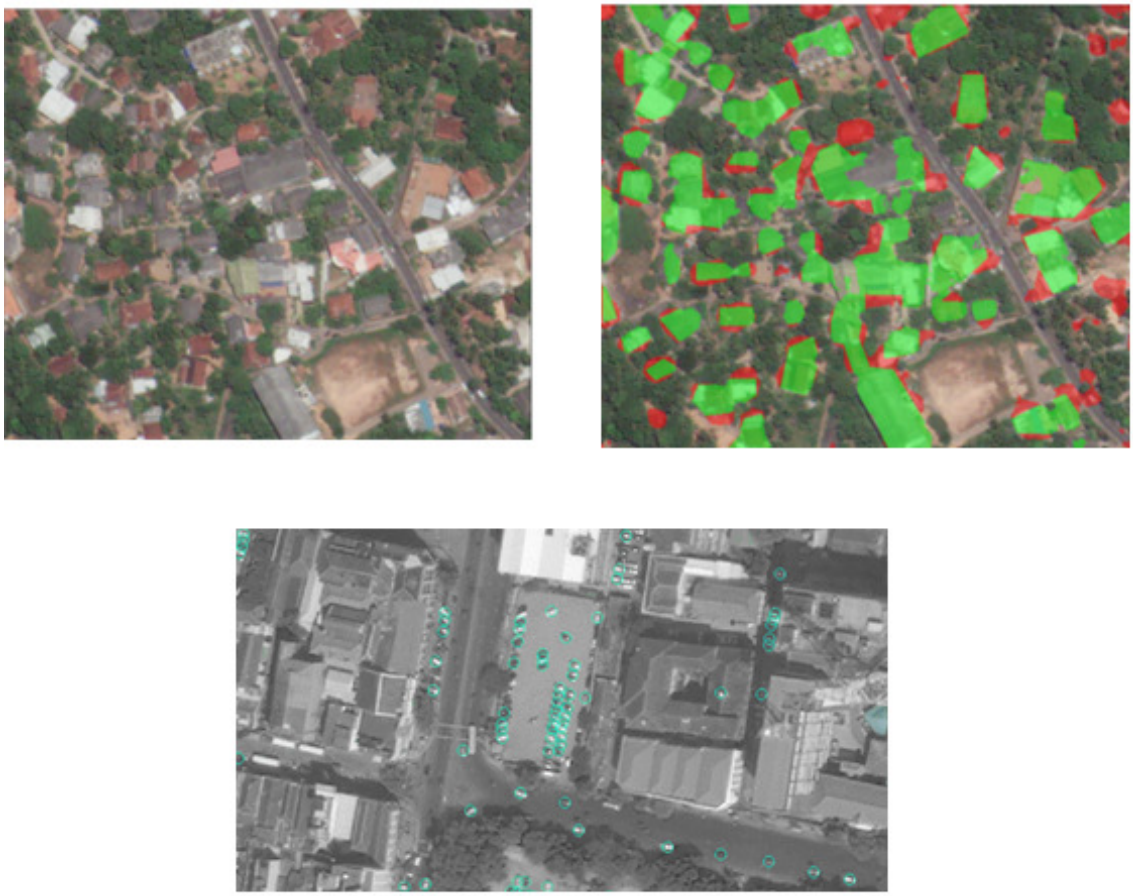
Finally, we multiply the inferred number of housing units in the village from Equation 6 by the average household size of that village in the sample, to obtain an estimate of the village-level population. This is then divided by the physical size of the village, derived from the boundary file, to obtain an estimate of the population density of each village in the sample.

This procedure generates estimates that correspond reasonably closely to the census. However, accuracy could be improved by obtaining direct counts of the number of housing units in each sample PSU, rather than indirectly obtaining estimates from the sample weights. This would eliminate sources of approximation error, such as dropping three percent of the villages that contain multiple PSUs. Therefore, the estimated accuracy of predictions using the HIES-based density measure can be interpreted as a lower bound. Because these approximation errors are largely uncorrelated with the satellite-based independent variables in the model, using the HIES-based density approximation as the dependent variable only slightly decreases the accuracy of the predictions reported in Table 9. This suggests that household survey weights, under the proper conditions, can be combined with readily available census-based population counts to obtain reasonable estimates of population density at fine geographic levels in surveyed areas.

B The 55 Divisional Secretariat divisions (sub-districts) from Engstrom et al. (2017)

Ambagamuwa, Ambalantota, Ambanpola, Bandaragama, Biyagama, Bulathsinhala, Colombo, Dehiwala, Devinuwara, Dodangoda, Doluwa, Dompe, Galle Four Gravets, Hali Ela, Hambantota, Homagama, Horana, Ingiriya, Kaduwela, Kalutara, Kamburupitiya, Katana, Kattankudy, Kelaniya, Kesbewa, Kirinda Puhulwella, Kolonnawa, Kotapola, Kotmale, Kurunegala, Madurawala, Maharagama, Malimbada, Manmunai North, Matara Four Gravets, Moratuwa, Nagoda, Negombo, Nuwara Eliya, Nuwaragam Palatha East, Padukka, Panadura, Panwila, Puttalam, Ratmalana, Rattota, Seethawaka, Sri Jayawardanapura Kotte, Thihagoda, Thimbirigasyaya, Tissamaharama, Udapalatha, Udunuwara, Ukuwela, and Uva-Paranagama.

Figure S1: Object Classification Examples



Note: Source: Engstrom et al. (2017). (Top Left) Raw satellite imagery. (Top Right) Classified image showing developed area building classifier. Areas in green show are true positive building classifications. Images in red are false positives. (Bottom) Example Car Classification. Cars classified by CNN are circled in blue.

Table S1: Publicly Available Population Products

Data Source	Resolution	Census information and other details	Distribution technique
LandScan (Oakridge National Laboratory, 2018)	30 arc seconds	LandScan uses annual mid-year national population estimates from the Geographic Studies Branch, US Bureau of Census (not the official census of the country).	LandScan distributes census counts in an area across grids based on the likelihood of being populated. This likelihood coefficient is calculated based on geospatial characteristics, including, roads, elevation, slope, and light intensity.
Facebook's High-Resolution Settlement Layer (HRSL) (Centre for International Earth Science Information Network, 2018)	1 arc-second (approximately 30m)	Most recent censuses. The population data have been developed for 18 countries: Algeria, Burkina Faso, Cambodia, Ghana, Haiti, Ivory Coast, Kenya, Madagascar, Malawi, Mexico, Mozambique, the Philippines, Puerto Rico, Rwanda, South Africa, Sri Lanka, Tanzania, Thailand, and Uganda.	High-resolution (0.5m) satellite imagery from DigitalGlobe. The satellite imagery were classified as settled or not based on computer vision techniques. Proportional allocation was used to distribute population data from subnational census data based on the settlement extents.
WorldPop 2010 and 2015 (WorldPop, 2018)	100 m	WorldPop 2015 uses the 2010 round of censuses; WorldPop 2010 uses the 2000 round of censuses	A semi-automated dasymetric modelling approach using random forest method incorporating geospatial datasets (e.g. settlement locations, settlement extents, land cover, roads, building maps, health facility locations, satellite nightlights, vegetation, topography, refugee camps).
Gridded Population of the World (GPW) v3 (2010) and v4 (2015) (Socioeconomic Data and Applications Center, 2018)	GPW v4: 30 arc second (approximately 1 km); GPW v3: 2.5 arc minute (about 5 km)	V4 uses the 2010 round of censuses, and V3 uses the 2000 round of censuses	Uniform weighting method at the GN (village) administrative level
Global Human Settlement (GHS) (European Commission, 2018)	250m	Residential population estimates from CIESIN GPWv4	Population estimates are disaggregated from census or administrative units to grid cells, informed by the Global Human Settlement Layer (GHSL) built-up area measures.

Table S2: Validation of Population Estimates at the Sub-district Level using Correlation Coefficient

Correlation	Census	Facebook	WorldPop 2010	WorldPop 2015	GPW 2010	GPW 2015	GHSL	LandScan
Census	1							
Facebook	0.999	1						
WorldPop 2010	0.948	0.870	1					
Worldpop 2015	0.999	0.997	0.870	1				
GPW 2010	0.997	0.991	0.865	0.993	1			
GPW 2015	0.997	0.994	0.877	0.994	0.995	1		
GHSL	0.997	0.998	0.875	0.997	0.991	0.995	1	
LandScan	0.126	0.116	0.162	0.125	0.151	0.132	0.122	1
Average discrepancy relative to census								
Persons	0	2468	15107	10360	4948	3663	2778	46936
Percent	0%	4.02%	21.03%	13.38%	6.72%	5.37%	4.79%	69.58%

Note: GPW=Gridded Population of the World; GHSL=Global Human Settlement Layer. Discrepancy is the absolute value of the difference between population estimates and the census at the sub-district level. Percent relative to the census is absolute discrepancy/census.

Table S3: Validation of Population Estimates at the Village Level using Correlation Coefficient

Correlation	Census	Facebook	WorldPop 2010	WorldPop 2015	GPW 2010	GPW 2015	GHSL	LandScan
Census	1							
Facebook	0.921	1						
WorldPop 2010	0.330	0.309	1					
Worldpop 2015	0.995	0.916	0.331	1				
GPW 2010	0.717	0.756	0.257	0.726	1			
GPW 2015	0.771	0.796	0.285	0.777	0.771	1		
GHSL	0.818	0.891	0.280	0.817	0.698	0.725	1	
LandScan	-0.028	-0.030	-0.021	-0.027	-0.017	-0.030	-0.029	1
Average discrepancy relative to census								
Persons	0	292	733	197	669	553	515	1684
Percent	0%	22.62%	72.12%	13.5%	51.02%	45.77%	42.42%	197.37%

Note: GPW=Gridded Population of the World; GHSL=Global Human Settlement Layer. Discrepancy is the absolute value of the difference between population estimates and the census at the village level. Percent relative to the census is absolute discrepancy/census.

Table S4: Comparing Built-up Area Measures at the Village Level (National sample)

R^2	Facebook	GUF	GUF+	GHSL
Facebook	1			
GUF	.807	1		
GUF+	.818	.862	1	
GHSL	.787	.761	.821	1
Average absolute discrepancy relative to Facebook	0	7.51	15.47	7.26

Note: GUF= Global Urben Footprint; GHSL=Global Human Settlement Layer.

Table S5: Comparing Built-up Area Measures at the Village Level between Engstrom et al. (2017) and Facebook (55 Sub-districts)

Measure	FB and Engstrom et al. (2017)
R^2	.794
Average discrepancy in built-up area	23.60%
%Villages with discrepancy greater than 5%	78.49%
%Villages with discrepancy greater than 10%	53.92%
%Villages with discrepancy greater than 25%	34.85%

Note: Discrepancy is the absolute difference between built-up area measures at the Village level. FB=Facebook. Engstrom et al. (2017) uses Digital Globe imagery to derive features and textures.

Table S6: Comparing Built-up Area Measures at the Village level using Correlation Coefficient (National sample)

Correlation	Facebook	GUF	GUF+	GHSL
Facebook	1			
GUF	0.898	1		
GUF+	0.904	0.929	1	
GHSL	0.887	0.874	0.912	1
Average absolute discrepancy relative to Facebook	0	7.51	15.47	7.26

Note: GUF= Global Urban Footprint; GHSL=Global Human Settlement Layer.

Table S7: Comparing Built-up Area Measures at the Village Level between Engstrom et al. (2017) and Facebook (55 Sub-districts) using Correlation Coefficient

Measure	FB and Engstrom et al. (2017)
Correlation	0.891
Average discrepancy in built-up area	22.52%
%Villages with difference greater than 5%	74.9%
%Villages with difference greater than 10%	51.37%
%Villages with difference greater than 25%	33.13%

Note: Discrepancy is the absolute difference between built-up area measures at the Village level. FB=Facebook. Engstrom et al. (2017) uses Digital Globe imagery to derive features and textures.

Table S8: Accuracy of Population Density Predictions, without Log Village Area as a Covariate

Imagery Type	In-sample R^2	Out-of-sample R^2	Number of variables		Number of GNs
			Candidate	Selected	
A. National Sample					
1. No satellite imagery	0.499	0.359	34	23	13970
2. Low resolution & public	0.605	0.468	34	30	13970
3. All Public	0.804	0.658	38	25	13437
B. 55 sub-districts sample					
1. No satellite imagery	0.422	0.361	34	3	1360
2. Low resolution & public	0.697	0.581	34	13	1360
3. All Public	0.803	0.658	38	5	1275
4. All Public + Propriety texture	0.837	0.728	45	10	1275
5. All	0.917	0.847	66	37	1178

Note: The results are based on a Poisson regression model whose dependent variable is census village population density, and the independent variables are selected based on LASSO regularization among the following: different types of satellite imagery sources as described in panel B in Table 1, district fixed effects, and a binary indicator for urban villages. The models with "No satellite imagery" use only the district fixed effects, an indicator for urban villages, and log village-area as independent variables. "Low-resolution public" indicators refer to the low-resolution and public indicators as defined in Table 1. "All public" indicators refers to all the public imagery based indicators, including the high-resolution imagery based indicators described in Table 1. "Proprietary texture" indicators include very high-resolution texture features based on Digital Globe from Engstrom et al. (2017) as defined in Table 1. "All" imagery refers to all public and proprietary texture indicators, and object identifiers, such as cars, roofs etc., as defined as in Table 1. The out-of sample R^2 is obtained from stata's crossfold command using five-fold cross-validation.

Table S9: Marginal Effects from Poisson Regression Results of Census Village Population Density on LASSO selected variables

	National Sample	55-sub district sample
Public indicators		
Night-time lights (March 2014), Maximum	3.006***	0.654
	(1.142)	(1.514)
Mean Slope		-67.73***
		(16.57)
Built-up area from GUF	5.749***	
	(2.228)	
Built-up area from GHSL	4.362***	
	(1.177)	
Built-up area from Facebook	12.22***	-3.281
	(2.657)	(5.707)
Built-up area from GUF+	7.231***	25.67***
	(1.526)	(3.756)
Proprietary indicators		
% of roads that are paved city (4 m width)		8.884***
		(2.914)
% of roads that are main paved (5 m width)		-9.340***
		(3.62)
% of roads that are minor paved (4 m width)		-0.422
		(4.537)
% shadow pixels covering valid area		50.06***
		(16.31)
NDVI (mean) scale 32		-964.1**
		(457.4)
Total built-up area		0.00406***
		(0.00117)
Fraction of total roofs that are clay		-35.95***
		(4.737)
Fraction of total roofs that are aluminum		-22.79***
		(8.555)
Fraction of total roofs are asbestos		-27.37***
		(7.247)
log number of cars		-251.5***
		(75.94)
% of Village agriculture that is paddy		3.907***
		(1.338)
Pantex (human settlements) mean, scale 8m		-173.9
		(283.6)
Histogram of Oriented Gradients (scale 16m), mean		-17.14**

Continued on next page

Table S9 – *Continued from previous page*

	National Sample	55-sub-districts sample
		(8.38)
Gabor filter (scale 64m, features 6), mean		-387.3
		(1,179)
Gabor filter (scale 64m, features 14), mean		1,927
		(1,409)
Geographical descriptors		
Ln Village area	-1,633***	-2,263***
	(82.59)	(264.2)
Binary for urban Villages	161.0**	
	(73.13)	
District Fixed Effects (base: Ampara)		
Colombo	376.3**	347.4**
	(168.8)	(150.2)
Gampaha	114.1***	
	(43.48)	
Kandy	166.2**	
	(71.53)	
Kurunegala	-159.9***	
	(48.97)	
Puttalam	-184.3**	
	(93.83)	
Trincomalee	504.0***	
	(96.7)	
Badulla		-336.2
		(321.8)
Observations	12,088	1,051

Note: Marginal effects from the Poisson regression model of census population density on variables selected from LASSO selection are reported. Standard errors clustered at the sub-district level in parentheses; *** implies $p < 0.01$, ** implies $p < 0.05$, and * implies $p < 0.1$. GPW=Gridded Population of the World; GHSL=Global Human Settlement Layer; GUF=Global Urban Footprint; NDVI=Normalized Difference Vegetation Index. The original list of variables before LASSO selection included all publicly available satellite indicators for the national sample, and all public and proprietary indicators for the 55 sub-district sample.

Table S10: Out-of-sample Accuracy of HIES-based Village Population Count Estimates, 55 Sub-Districts

	R^2	SRC	Mean AE	Median RE	Median RE
A. Model-based estimates					
HIES-based	0.26	0.51	642	42%	30%
HIES-based (with balance correction)	0.24	0.49	641	42%	31%
Census-based	0.28	0.36	694	58%	32%
B. Existing areal interpolation estimates					
WorldPop 2015	0.30	0.53	232	13%	13%
Facebook HSRL 2015	0.25	0.47	316	20%	15%
GHSL Grid 2014	0.19	0.42	563	39%	28%
GPW 2015	0.54	0.51	729	44%	28%
GPW 2010	0.05	-0.01	901	52%	37%
WorldPop 2010	0.02	0.22	946	64%	34%
LandScan 2010	0.00	-0.19	1967	137%	90%

Note: Note: HIES=the Household Income and Expenditure Survey; GPW=Gridded Population of the World; GHSL=Global Human Settlement Layer; GUF= Global Urban Footprint; AE= Absolute; RE=Relative Error; SRC=Spearman Rank Correlation. We retain only the villages in the 55 sub-districts for which the proprietary data are available. The model using population count as the dependent variable is estimated on sub-sample of 414 villages covered in 2012-2013 HIES survey. The satellite indicators used as covariates include all publicly available and proprietary indicators. Out-of-sample predictions for testing the accuracy with the census are conducted on 946 villages not covered in the HIES. The weights for the HIES-based model are $\frac{1}{VillagePopulation}$. Balance correction refers to adjustment of weights based on the predicted probability of the village being sampled in HIES ($\widehat{INHIES}$) based on a probit model with village sampled in HIES as the dependent variable and all LASSO-selected variables as independent variables. The balanced corrected weights are $\frac{1}{VillagePopulation X \widehat{INHIES}}$.