INFORMATION TECHNOLOGY FOR DEVELOPMENT
2021, VOL. 27, NO. 2, 263–292
https://doi.org/10.1080/02681102.2020.1811945

Routledge
Taylor & Francis Group

Open data for algorithms: mapping poverty in Belize using open satellite derived features and machine learning*

Jonathan Hersh [a], Ryan Engstrom[b] and Michael Mann[b]

[a]Argyros School of Business, Chapman University, Orange, CA, USA; [b]Department of Geography, George Washington University, Washington DC, USA

ABSTRACT

Several methods have been proposed for using satellite imagery to model poverty. These include poverty mapping using convolutional neural networks applied either directly or using transfer learning to high resolution satellite images, or combinations of methods that combine satellite imagery with standard methods. However, these methods require proprietary imagery which, given their cost and infrequent acquisition, may render these advances impractical for most applications. The authors investigate how satellite-derived poverty maps may improve when incorporating features derived from Sentinel-2 and MODIS imagery, which are both open-source and freely and readily available. The authors estimate a poverty map for Belize which incorporates spatial and time series features derived from these sensors, with and without survey derived variables. They document an 8% percent improvement in model performance when including these satellite features and conclude by arguing that Open Data for Development should include open data pipelines where possible.

KEYWORDS

Poverty; small area estimation; satellite imagery; remote sensing; open data

1. Introduction

The Open Data for Development (OD4D) movement promotes the publication of government data and statistics, under the belief that increased government transparency is crucial for promoting economic growth. Since its inception, the global partnership has sponsored stakeholders such as government and multilateral development banks to release thousands of datasets, promoted data literacy, tracked data on impact monitoring, and promoted the innovative use of data for development. OD4D has been cited as an important tool in poverty reduction, corruption prevention, and in holding accountable politicians and other elected officials (Maail, 2017; World Bank Group, 2017). What has

CONTACT Jonathan Hersh hersh@chapman.edu

*We thank Emmanuel Abuelafia, Alejandra Mejia Castellanos, Guillermo Lagarda Cuevas, Lucia Martin Rivero, Cassandra Rogers, and Marta Ruiz Arranz at the Inter-American Development Bank for guidance and assistance during this project. We thank Lesley Cruz, Diana Castillo-Trejo, and Francine Gongora at the Statistical Institute of Belize for assistance with the Census and LFS survey data, and Maritza Canto for assistance with the Enumeration District shapefiles. Special thanks to Drs. Leopold L. Perriot and Geraldo Flowers for sharing the Census and LFS data with our research team, and to Mark Antrobus at the Ministry of Human Development for feedback on the model. Presentations and feedback at the Inter-American Development Bank, the Statistical Institute of Belize, and the Ministry of Human Development of Belize greatly improved our project. This article was developed with the support of the Inter-American Development Bank. However, the opinions expressed in this article are those of its authors and do not necessarily reflect the point of view of the Inter-American Development Bank, its Board of Directors, or the countries it represents. Two anonymous referees, and the Review Editor Dr. Pamela Abbott, helped immensely in producing the final article.
Production Editor: Pamela Abbott is the accepting Editor for this manuscript.
This article has been republished with minor changes. These changes do not impact the academic content of the article.

received relatively limited attention are the data pipelines that are used to produce open data, and whether these inputs themselves are in fact open and transparent.

In this paper, we evaluate the use of proprietary satellite imagery for the purposes of poverty measurement, and test an alternative source of open-source satellite imagery that is freely and readily available. Poverty eradication is the first of the UN Sustainable Development Goals (SDGs). However, 'data gaps' in the coverage of poverty statistic is persistent, despite international efforts to increase their development. As many as 57 countries produced one or fewer data points on poverty in the decade between 2002 and 2011 (Serajuddin et al., 2015). Producing sub-national poverty estimates at regular frequencies is burdened by the expense of conducting reliable consumption surveys. Kilic et al. (2017), estimate an average direct survey cost of $105 per household surveyed, and technical assistance costs on average are $613,000 per survey in Latin America. Combined, total costs are approximately $2 million per survey. Despite the innovations of rapid poverty assessment approaches such as Pape and Mistiaen (2018) and SWIFT (Yoshida et al., 2015), conducting reliable poverty surveys with sufficient frequency remains prohibitively expensive in most countries.

Given finite budgets and limited technical capacity to mount these surveys, several researchers have proposed using Big Data to assist in the generation of sub-national estimates of poverty, such as metadata from cellular phones (Blumenstock et al., 2015) or satellite imagery (Engstrom et al., 2017; Jean et al., 2016). Most of these methods, however, rely on the use of expensive and proprietary cellular metadata, or high spatial resolution satellite imagery. For the hypothetical statistical agency that cannot commit $2 million to mount a survey, the proposition of purchasing expensive data such as high-resolution satellite imagery amounts to a sick patient considering two medicines, the expense of either of which is too dear to bear.

This paper investigates the extent to which open-source, Big Data can be meshed with existing survey data to alleviate the lack of frequent sub-national poverty estimates. Using Belize as a test case, we utilize freely available, open-source satellite imagery to build sub-national estimates of income poverty, and determine the extent to which features from satellite imagery act as substitutes and complements to survey-based estimates of poverty. Using two waves from the 2017 Labor Force Survey (LFS) in Belize, we estimate machine learning models to predict household labor income as a function of survey and satellite variables. Features generated from satellite imagery are derived to capture both cross-sectional (spatial) information as well as time-series properties. This information may be able to capture, for example, drought conditions in remote agricultural areas otherwise unobserved in surveys. We utilize four machine learning models – Ridge Regression, Elastic Net Regression, Random Forest, and Extreme Gradient Boosted Trees – to predict household income from satellite and survey characteristics. Despite substantial model agreement, we control for model uncertainty by creating 'ensemble' estimates of household income. Additionally, we create an ensemble estimate of poverty rates using information from all four estimated models, which may be more robust to model uncertainty than a single poverty model.

We find that household-level income models used to generate Enumeration District (ED) poverty rates improve when incorporating satellite variables. Satellite and survey models explain 53.7% of the variation in average incomes between predicted and true average ED income, compared to 49.7% of the variation using survey data alone, an improvement of 8%. Satellite models alone explain 29.8% of the variation between predicted average ED income and true average ED income. Altogether, these are not stunning arguments for the use of open-source Big Data. However, we find that models improve most significantly for the poorest households. Given that much of the poverty in Belize occurs in rural areas, we believe satellite variables capture important features of income that are not observed by surveys.

Open-source and freely available satellite images may hold many potential benefits for resource constrained agencies. For one, statistical agencies can commit to the price of $0 for open-source

imagery in perpetuity. In comparison, a statistical agency that incorporates proprietary data into their statistical pipeline opens themselves to price gouging as proprietary data providers have pricing power due to 'lock-in' type effects (Arthur, 1989). It is possible that even with competition among data providers, any surplus from using Big Data at statistical agencies may eventually be captured by proprietary data providers because of lock-in effects due to the difficulty of moving from established data pipelines. Thus, it is crucial to consider open-source alternatives to proprietary data providers.

Previous research has explored using a variety of approaches of open source Big Data to estimate poverty including (Jean et al., 2016; Pokhriyal & Jacques, 2017; Steele et al. 2017). In a number of these studies they use deep learning convolutional neural networks on Google Static Maps to map poverty (Jean et al., 2016). While the imagery is available and is of high spatial resolution, the date of the imagery used to create the static maps is highly variable and difficult to determine. Additionally, these data are limited to only the visible bands, Blue, Green and Red. Hence, models built using this data are limited to using only the visible bands, and may exclude useful methods to estimate vegetation that rely on bands outside of the visible spectrum. Other research uses call data records (CDRs) that are often provided by private cell phone companies (Pokhriyal & Jacques, 2017, Steele et al. 2017). While these data can be used to estimate poverty, access to these data are limited as they are collected by private companies. It often costs tens of thousands of dollars to access this data. Additionally, there is some concern whether CDR data is representative of the entire population (Jacques, 2018). In this study we use only completely open and readily updated (taken at least every 5 days) satellite data combined with survey and census data that are typically available to in country statistical agencies. Therefore, this paper fills a necessary gap in the literature whereby we explore whether these open-source satellite data alternatives may be of use to the prototypical statistical agency who does not want to rely on proprietary and costly satellite data.[1]

2. Data and background

2.1. Country context

Belize is a small, Caribbean country located in Central America. A former British colony, previously known as British Honduras, Belize became independent in 1981 and has since grown to roughly 400,000 inhabitants. Belize is classified as an upper-middle income country, with a GDP per capita of $4,905 USD in 2017. Its economy is heavily dependent on tourism, including roughly a million visitors who arrive via cruise ships and half a million which arrive and stay overnight. Its industry is heavily based in agriculture, with the majority of Belize's exports being raw and processed agriculture including sugar, bananas, frozen fish, rubber, tobacco and petroleum. Despite having a small footprint, the country has a low population density of 17.2 people per square kilometers. This is immediately evident in the population density map in Figure 1, showing that outside of the main cities there is ample brush and vegetation. Remote sensing may play a critical role here in capturing measures of agricultural output that may be absent from surveys.

Belize is an exemplar of a country for which open-source Big Data methods may be a boon. The country last produced a poverty assessment in 2009 (Belize National Human Development Advisory Committee, 2010). Since then, no sub-national poverty statistics have been produced to our knowledge. The Statistical Institute of Belize (SIB) conducts bi-annual Labor Force Surveys with sample sizes that allow for models of small area estimation to be estimated. We hypothesize that countries most likely to benefit from methods similar to the one in this paper are those that have a substantial rural population, perform frequently auxiliary surveys, such as Labor Force Surveys, but do not have frequent consumption surveys, and who have a desire to upgrade statistical capacity around these tools.

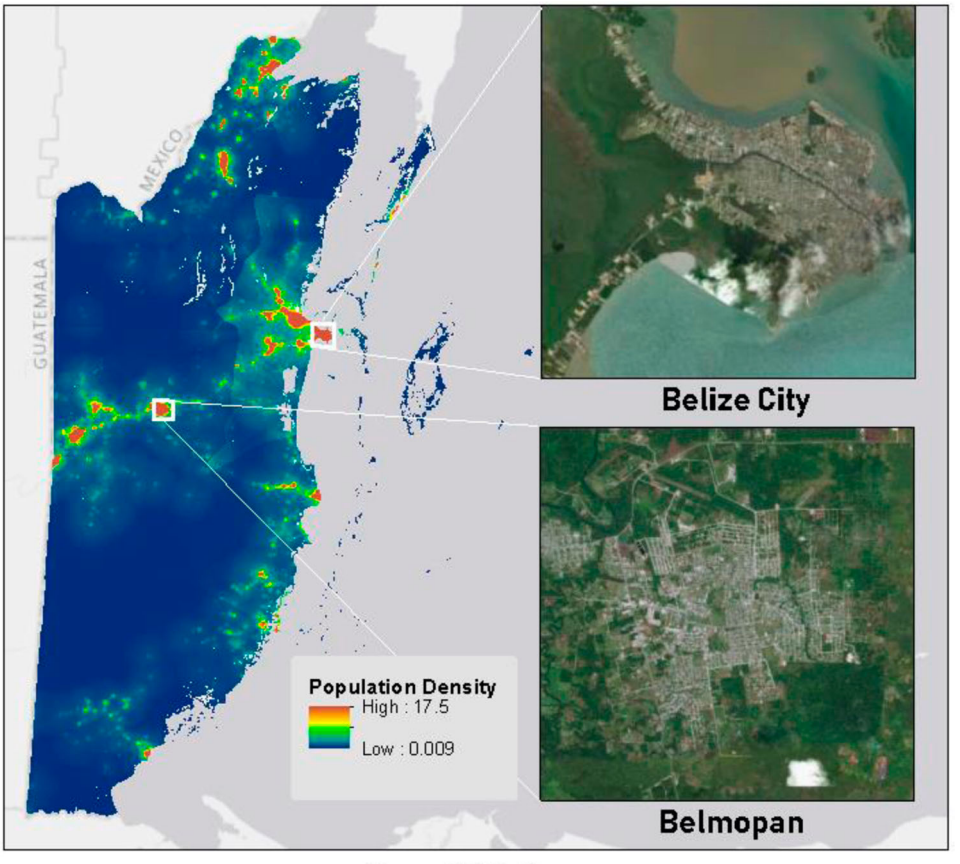


Figure 1. Belize Population Density (persons per ~100 m pixel).

2.2. Survey data

In order to generate estimates of household income we utilize two datasets: first, a labor force survey that asks critical questions about income but has a relatively small sample size, and second, the national census, which covers all households but does not directly ask about household income.

2.2.1. Labor Force survey and census

We derive household income statistics from the April 2017 and September 2017 waves of the Belize Labor Force Survey. The April wave surveyed 2,331 unique households and the September wave surveyed 2,320 households, which were repeated cross-sections and not necessarily the same sampled households. This resulted in 4,651 unique households across the two waves. After removing missing values, this resulted in a data set of 3,658 household level observations. We split the observations into 75% training data, which will be used to estimate our household level model, and 25% testing data, which will be used to validate our model at the household and enumeration district level. This testing and training split of the data is a standard method in machine learning (James et al., 2013), which is a necessary step to prevent overestimation of model performance. However, this technique of splitting data into testing and training sets is infrequently used in generating poverty maps using survey data alone. In comparing our model performance to other models, the most directly comparable statistics will be our 'in-sample' performance metrics, which still reflect cross-validation as described in the

following section. While reserving 25% for the test set may be large for some applications, we believe this is appropriate to capture test set performance metrics with sufficient accuracy.

We derive 37 co-variates from both the Census and Labor Force Surveys, which derive from identical survey questions. While the set of co-variates may seem small, consider that a popular software for building poverty maps provided by the World Bank 'PovMap[2]' can estimate at most 25 variables. To these survey data we add numerous satellite variables which we derive from openly available satellite imagery.

2.3. Remotely sensed data

In order to summarize satellite images at some geographic level, we process images to create summary statistics that capture different spatial and temporal aspects of the imagery. These statistics will then provide information that correlates with conditions on the ground, such as indicators like building and vegetation patterns that are correlated with poverty. Satellite data provides several clues about the conditions of households on the ground. Images might be able to provide information about the size of homes, the types and intensity of land-use, or even indicators of successful and poor agricultural seasons. The challenge then is to extract a set of metrics from imagery that can help describe some of these attributes.

In this study we explore two approaches. The first is a method known as Contextual Features, which looks at spatial and spectral patterns within neighborhoods within a single time period. Contextual Features can help us understand texture, orientation, complexity, and continuity of neighborhoods, or groups of pixels. The second approach looks at time series features, examining the change of each pixel in an image over time. Here we can extract features like maximums, means, trends, and sudden shifts, for a variety of metrics including rainfall and greenness. A summary of the sensor and their characteristics are outlined in Table 1 below and discussed in more detail in the following sections.

2.3.1. Sentinel-2 imagery: contextual features

In the creation of Contextual Features, we use imagery from the Sentinel 2 sensors (both A and B). Sentinel 2 imagery measures reflected energy from the sun in 12 wavelengths from the visible bands (Blue, Green and Red) into the Near Infrared (NIR) and Short-Wave Infrared (SWIR). Recall that a visible color image is composed of different visible 'channels' or 'bands,' which each capture different wavelengths, typically Blue, Green and Red bands. We focused on the visible (Blue, Green, Red) and NIR bands of Sentinel 2 because they have the highest spatial resolution with a pixel size of 10 m. One of the difficulties in working with these types of data in a country such as Belize is cloud cover, because the sensor measures reflected sunlight. In order to overcome these issues, we use Google Earth Engine to create a cloud-free mosaic of the entire country by selecting a cloud free pixel for each location over a period of time. This was done by selecting the median pixel in each band from January 1, 2017 to March 31, 2018 from every image that was collected by the

Table 1. Description of satellite sensor, application, and remotely sensed data.

Satellite	Application	Resolution	Period
MODIS	NDVI time series	250m spatial resolution 16-day temporal	January 1, 2013 to December 31, 2017
CHIRPS	Precipitation time series	0.05 degrees spatial resolution Dekad temporal	January 1, 2013 to December 31, 2017
Sentinel	Contextual features	10m spatial resolution Annual median temporal	January 1, 2017 to March 31, 2018
PALSAR	HH HV properties	25m spatial resolution Annual median temporal	January 1, 2017 to March 31, 2018

Sentinel 2 sensors (images are acquired every 5 days). These data provide the spatial detail required to observe spatial patterns across the landscape.

2.3.2. MODIS imagery: time series data

MODIS sensors acquire images for much of the world twice daily. As such, while MODIS has a low spatial resolution (250 m), it is compensated by its high temporal resolution – that is it has a high revisit rate. Thereby, these data provide rich information for time series statistics. Moreover, because MODIS has been in orbit for many years, a longer time series is available that allows us to summarize the five years leading up to the study date from January 1, 2013 to December 31, 2017.

We calculate the normalized difference vegetation index (NDVI) from MODIS. NDVI is commonly used to monitor the status of crops, forests, and ecosystems. NDVI is sensitive to the amount of chlorophyll in any location and is used to observe approximate levels of plant productivity and health. Given the relatively small scale of agriculture in Belize, we derive the NDVI using the 250 m vegetation products from the MODIS sensors.

2.4. CHIRPS rainfall: time series data

We also examine the time series properties of rainfall as measured by the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS). CHIRPS is a 30+ year old quasi-global rainfall dataset. CHIRPS incorporates 0.05° resolution satellite imagery with in-situ station data to create gridded rainfall time series for trend analysis and seasonal drought monitoring. In this case we resample the rain data to 75 m spatial resolution to ensure that each enumeration area has an observation associated with it. We collect precipitation by dekad (Funk et al., 2014). There are three dekads in a month, the first two being 10 days long, and the third being the remaining days in the month. Because CHIRPS data has a similar high frequency and availability as MODIS data above, we provide the denser set of summary statistics outlined in Table 3 for low-spatial resolution data.

2.5. Synthetic aperture radar

We utilize the Japanese sensor PALSAR/PALSAR-2 mosaic data to provide 25 m resolution synthetic aperture radar (SAR) data. SAR can be used to create three-dimensional reconstruction of objects, such as mountains and landscapes (Kirscht, 1998; Kirscht & Rinke, 1998). The PALSAR sensor has two polarizations HH and HV. HH measures the proportion of horizontally transmitted waves which return horizontally to the sensor. HV measures the part of the emitted waves which are polarized at the earth's surface and return vertically to the sensor. HV and HH are sensitive to the physical properties of objects on the ground including vegetation, and urban environments amongst others.

Combined SAR, HH, and HV polarization data can provide critical information on the structure of objects on the ground. SAR has been successfully used to create global forest/non-forest maps (Shimada et al., 2014), for assisting in remote crop classification (McNairn et al., 2009), to mapping flooding events (Shan et al., 2010).

3. Methods

3.1. Machine learning small area estimation

Most surveys which measure income or consumption do not sample all areas where policy makers would like to have estimates of poverty or welfare. Many low-income areas are largely inaccessible or sparsely populated. Even when they do sample all areas, there may not be sufficient observations to generate welfare statistics at a spatial resolution. As a result, most estimates of welfare at the local level are generated through small area estimation, techniques which typically match a target survey, which measures the variable of interest (poverty, consumption or income), and a census, which

contains sufficient observations from which one can accurately calculate welfare. This approach requires a model of household-level welfare, $y_{h,c}$, which is the income or consumption level of household h measured in local area c. One approach is to assume a (log) linear relationship between household characteristics $x_{h,c,k}$ and income/consumption, which takes the form:

$$\ln y_{h,c} = \sum_{k=1}^{K} x_{h,c,k} \cdot \beta_k + \epsilon_{h,c} \tag{1}$$

where $\epsilon_{h,c} \sim N(0, \sigma)$ is the unexplainable error term. Note that if we build a sufficiently accurate model of income or consumption, and the true income process is linear, $\widehat{y_{h,c}} = y_{h,c} = \sum_{k=1}^{K} x_{h,c,k}\beta_k$. If we restrict our household characteristics $x_{h,c}$ to those that are also available in the census, and we have recovered a model of household income or consumption that remains sufficiently accurate in the survey and census, we can apply the parameter estimates obtained from the census $\widehat{\beta_k}$ to derive $\widehat{y_{h,c}}$ for every household in the census. This provides a method to compute welfare for each household in the census. The resulting estimate of welfare is:

$$\ln \overline{\widehat{y_c}} = \frac{1}{n_c} \sum_{h \in c} \ln \hat{y}_{h,c} = \sum_{k=1}^{K} x_{h,c,k} \cdot \widehat{\beta_k} \tag{2}$$

where n_c is the number of households or population in cluster c, and $\ln \overline{\widehat{y_c}}$ is the average welfare statistic of interest in the cluster.

Several problems may arise. For one, the true relationship between welfare $y_{h,c}$ and household characteristics $x_{h,c}$ is likely non-linear and difficult to model via ordinary linear regression (OLS) (Afzal et al., 2015). A common refinement of this method is to use simulation methods to improve the asymptotic properties of $\overline{\widehat{y_c}}$ (Elbers et al., 2003). We take a different approach, and instead use machine learning to model $\widehat{y_{h,c}}$ under the logic that machine learning will recover a better household-level model of welfare. In particular, we aim to better capture the likely non-linear relationship between satellite and census features and income. We estimate models of the form:

$$\ln y_{h,c} = f(X_{h,c}) + \epsilon_{h,c} \tag{3}$$

where $f(.)$ is estimated using four separate machine learning models. We estimate the following four models: (1) Ridge Regression (Hoerl & Kennard, 1970), (2) Elastic Net regression (Zou & Hastie, 2005), (3) Random Forest (Breiman, 2001), and (4) Extreme Gradient Boosted Trees (Friedman, 2001). We lastly estimate model (5) Combination, which creates a simple average of the four estimated models. This last model is a simple ensemble of several models that may retain the separate strengths of each model.

A detailed discussion of the methods is beyond the scope of this article.[3] However, one important note is that the first two methods are linear models that use machine learning for variable selection. The second two are tree-based ensemble methods of many regression trees. This is important in that linear models may perform worse in extrapolation if the support into which they are predicting are sufficiently different from the areas in which they have been estimated. For these reasons, tree models (models 3–4) may be more robust in their predictions.

Another concern for our modeling process is the likely spatial autocorrelation present in the units of analysis, enumeration districts. Generally, random sampling of spatially autocorrelated units will overstate predictive performance (Bahn & McGill, 2013). This becomes even more pronounced when analyzing panel or repeated cross-sectional data (Groger et al., 2020). One proposed solution is spatially stratified cross-validation, in which, during the process of cross-validation, units are sampled from distinct geographic areas (Pokhriyal & Jacques, 2017). In traditional 10-fold cross-validation, data are split into 10 distinct folds. A model is estimated using data from nine of the folds, and that model is used to predict into the 10th, or withheld, fold. The process is repeated until all folds

have predicted values. In spatially stratified cross-validation, data are distributed into folds according to distinct geographic units. For the process of training our models we employ 5-fold cross-validation at the Enumeration District level. That is, during the training process, we partition data into five folds, such that households from the same ED are assigned to the same fold.[4] Hyper-parameters are chosen according to those that minimize root mean squared error (RMSE) as calculated according to the spatially stratified cross-validation outlined. Our 25% test sample is taken from a random sample of enumeration districts thus all districts are represented, although none of these enumeration districts are used during any step of the training process.

3.2. Satellite imagery methods overview

Two primary methods are used to extract useful information from raw remotely sensed imagery, contextual features and time-series features. Raw images of a location, while useful for human interpretation, provides little useful information to a computer. Instead we must devise ways to extract specific components of an image that might be useful. For instance, does the image have many long lines pointing in the same direction, or are the lines short and complex? This type of information will be captured by 'Contextual Features' in the following section. Alternatively, we can look at a series of images over time and see if there were any sudden shifts in rainfall or greenness. This type of information will be described in 'Time-Series Features' in a later section.

Readers may be curious how this approach differs from the use of Convolutional Neural Networks (CNN), a popular method for computer vision in computer science (LeCun et al., 1998). CNNs apply a series of filters over an image to produce a series of feature maps, which highlight certain aspects of the images as determined by the specific filter. With CNNs, the parameters of the filters are optimized through model training, which requires thousands or millions of example images. Our method here also uses filters applied to images, only our filters are specifically purposed for extracting information from satellites. In both approaches the outputs from the feature maps are applied to statistical models to associate the feature maps to predict poverty or income of an area. The relative merits of using a CNN versus intermediate features have been discussed in Engstrom et al. (2017) and we refer the reader there for more detail.

3.3. Contextual features

One powerful set of methods to summarize satellite images is known as contextual features. Contextual features are information that represent the spatial and spectral values derived from satellite imagery based on neighborhoods or groups of pixels. In the past we have shown that these features are strongly correlated with population and poverty variation within Sri Lanka and Ghana (Engstrom et al., n.d.). For the most part, past research of this nature has used very high spatial resolution imagery (2 m spatial resolution and lower). While these data provide a tremendous amount of detail, there are major drawbacks including high cost, and difficulty covering large areas. Recent research has used data from the freely available, Sentinel-2 sensors, which have extensive spatial coverage, thus allowing us to easily and freely, collect imagery over large areas such as entire countries (Pesaresi et al., 2016; Verrelst et al., 2012).

A cloud free image, Sentinel-2 mosaic was used as the input to calculate contextual features using the Python package SpFeas. SpFeas is an open-source Python library for processing contextual image features from satellite imagery. The 11 contextual features calculated are as follows (Table 2):

Contextual features are created by comparing central pixels with their neighbors and then reporting this value back to the central pixel – 10 m in this case. Thus, contextual features measure the 'context' in which an individual pixel is situated, using information from surrounding pixels in addition to the central pixel's value itself. The number of neighboring pixels considered in the comparison is the scale, which varies by the feature being calculated.

When applied to satellite imagery, the features capture 'texture' and spectral values of neighborhoods. As an example, the Pantex feature captures the minimum contrast between a pixel and its

Table 2. Description of contextual features used in the analysis.

Name	Description	Interpretation	Source
Gabor Filter	A linear Gaussian filter used for edge detection	Finds edges of buildings and determines if they are in similar directions.	(Mehrotra et al., 1992)
Histogram of Oriented Gradients (HOG)	Captures the orientation and magnitude of the shades of the image	Finds the orientation of edges of buildings and groups them together.	(Dalal & Triggs, 2005)
Lacunarity (LAC)	Describes the extent of gaps and holes in a texture.	Finds gaps within areas. Can determine if buildings are close together or have space between.	(Myint et al., 2006)
Local Binary Patterns Moments (LBPM)	Define contiguous regions of pixel groups and sorts them into a histogram	Finds buildings and neighborhoods of different sizes.	(Ojala et al., 2002)
Line Support Regions (LSR)	Characterize line attributes	Characterizes the lengths of lines, typically roads and building edges.	(Ünsalan & Boyer 2005)
Normalized Difference Vegetation Index (NDVI)	The most widely used vegetation index that provides information about the health and amount of vegetation	Determines the presence or absence of vegetation.	(C J Tucker, 1979)
Oriented FAST and Rotated Brief (ORB)	Selects key points for image matching and object recognition. It is similar Speeded Up Robust Features (SURF).	Finds bright things such as buildings in imagery.	(Rublee et al., 2011)
PanTex	Is a built-up presence index derived from the grey-level co-occurrence matrix	Used to determine if areas have buildings or not. If buildings present, can help understand size.	(Pesaresi et al., 2008)
Structural Feature Sets (SFS)	Statistical measures to extract the structural features of direction lines	Finds road and building edges and characterizes the size and length.	(Huang et al., 2007)
Fourier Transform	Detects high or low frequency of lines	Can be used to determine if neighborhoods are on a grid pattern.	
Mean	The average brightness in the Blue, Green, and NIR bands	Finds bright and dark areas. Can help find vegetation.	

neighbors. Highly built-up neighborhoods tend to have greater contrast in all directions, which will create high values of this feature. In contrast, in rural areas the pixel's brightness will likely be similar to a neighbor in at least one direction, which will create a low minimum contrast. To help visualize what contextual features capture we present a number of features for Belize City in Figure 2 below.

For more detail on Contextual Features please refer to the Appendix.

3.4. Time-series features

To complement the contextual features described above, we also calculate several time series properties. Time series properties can play an important role in predicting household well-being. For instance, if an agricultural community has experienced below average rainfall for the last five years, this can be determined by looking at the time series for precipitation. Moreover, a variety of statistics can provide invaluable information, for instance, the maximum greenness of an agricultural area is correlated with agricultural yields and plant productivity (Mann et al., 2019; Mann & Warner, 2017). Time series can also pick up on the effects of drought, flooding, or even the slow sustained loss in productivity.

3.4.1. Dense temporal feature extraction

TS-raster (TS) is a python package for analyzing time-series characteristics from raster data. It allows feature extraction, dimension reduction, and applications of machine learning techniques for geospatial data and is available on GitHub.

TS's primary significance is the ability to provide an extensive set of time-series properties, including simple metrics like minimums or maximums, but also more complex ones like the number of peaks observed within a year, or the number of observations above or below the mean. TS should be able to meaningfully characterize the time series of high frequency data products like those from MODIS or CHIRPS. For a visual example of what kinds of properties TS extracts see Figure 3.

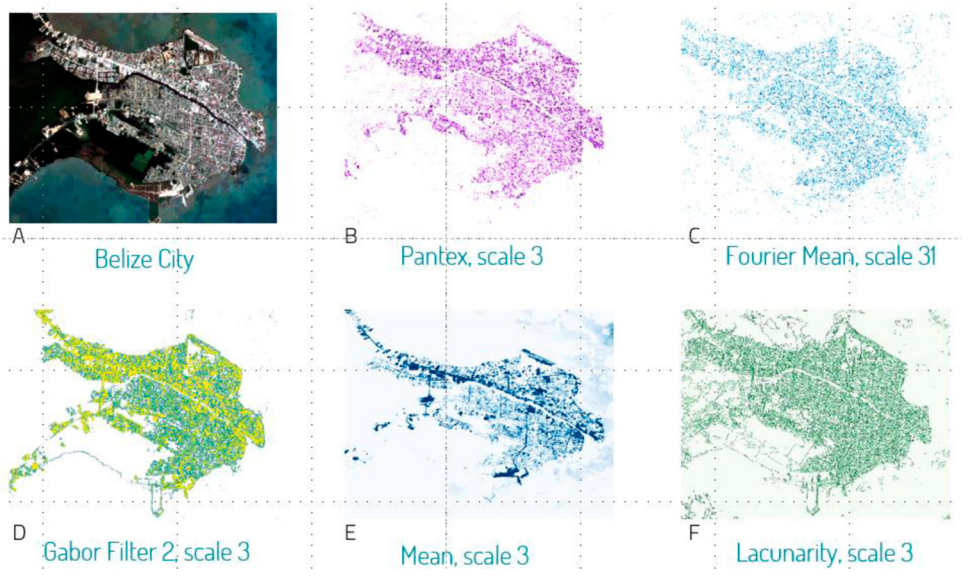


Figure 2. Belize City contextual features.

In this example the light grey line is a plot of a time series for a single pixel in an image. Purple boxes are used to highlight a series of time-series attributes that can be extracted with TS-Raster.

A summary of time-series attributes is provided below (Tables 3 and 4). The feature name indicates the naming convention used for data storage, the description provides a simplified description of that statistic, and use descriptions provide some context for how that attribute might be useful in our modeling. Table 3 provides a comprehensive list of statistics collected from data with very high temporal resolution (MODIS, CHIRPS). More summary statistics can be provided for this data because the time series has more observations, and is therefore more complex. Table 4, provides a full list of statistics gathered from sensors that are collected less often (ranging from once every 5–30 days depending on cloud cover). This is temporal data from sensors such as Sentinel-2 and PALSAR/PALSAR-2.

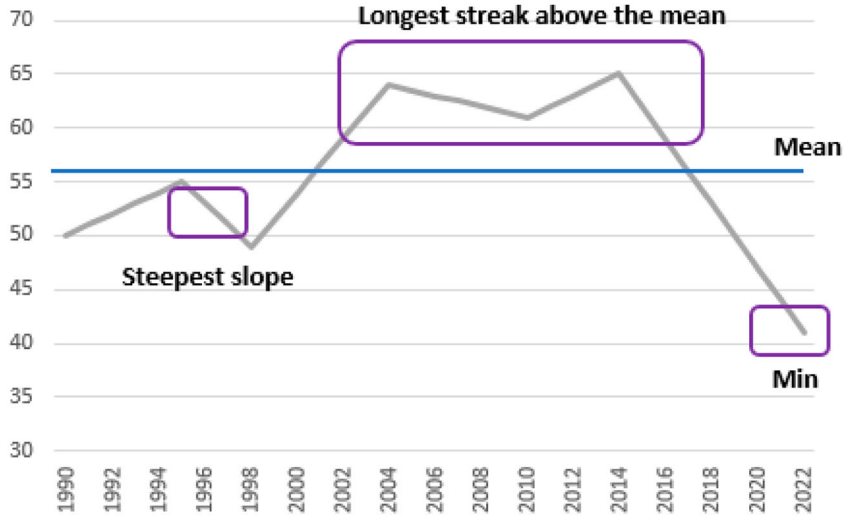


Figure 3. Examples of the time series extraction characteristics of the python package *TS-Raster*.

Table 3. Description of high temporal resolution time-series feature.

Name	Description	Interpretation
agg_linear_trend_*f_agg*'max'__ chunk_len_6_*attr*'slope'	Maximum observed trend during any 6 periods	Sudden positive shocks (flood)
agg_linear_trend_*f_agg*'min'__ chunk_len_6_*attr*'slope'	Minimum observed trend during any 6 periods	Sudden negative shocks (drought, land use change)
count_above_mean	Number of observations above the global mean	Persistent shifts up (increase in rainfall)
count_below_mean	Number of observations below the global mean	Persistent shifts down (decrease in rainfall)
last_location_of_maximum	Location of the periods maximum value	Time since maximum observed value (declining productivity)
last_location_of_minimum	Location of the periods minimum value	Time since minimum observed value (increasing productivity)
longest_strike_above_mean	Longest period of values observed above the global mean	Duration of persistent shifts up (flooding)
longest_strike_below_mean	Longest period of values observed below the global mean	Duration of persistent shifts down (drought)
maximum	Global maximum value	Highest observed greenness / rainfall
mean	Global mean value	Average observed greenness / rainfall
mean_change	Average change between any two periods in series	Instability in time series (irregular rain)
median	Global median value	Average observed greenness / rainfall
minimum	Global minimum value	Minimum observed greenness / rainfall
number_cwt_peaks__n_12	The highest number of peaks that occur in 12 periods	Number of crop rotations, unstable rainfall
number_cwt_peaks__n_6	The highest number of peaks that occur in 6 periods	Number of crop rotations, unstable rainfall
quantile__q_0.05	Value of the 5th percentile	Minima correcting for outliers
quantile__q_0.15	Value of the 15th percentile	Minima correcting for outliers
quantile__q_0.85	Value of the 85th percentile	Maxima correcting for outliers
quantile__q_0.95	Value of the 95th percentile	Maxima correcting for outliers
ratio_beyond_r_sigma__r_2	Ratio of values that are more than 2*std (x) away from the mean	Frequency of extreme values, flooding, shocks
skewness	Ratio of values that are more than 3*std (x) away from the mean	Frequency of very extreme values, flooding, shocks
sum_values	Sample skewness of x (calculated with the adjusted G1 coefficient)	Changes in distribution over time

Table 4. Description of low temporal resolution time-series features.

Name	Description	Interpretation
Med/Mn	Global median/mean NDVI	Average greenness, vegetation productivity
Min/Max	Global minimum/maximum NDVI	Max/Min vegetation productivity
P5/P25/P75/P95	5th, 25th, 75th, 95th percentile value NDVI	Min(Max)imal measures of greenness robust to outliers such as clouds
Sum	Sum of all NDVI values	Persistence of vegetation and productivity
Std	Standard deviation of NDVI values	Stability of greenness and productivity
LS_distr_mag_2012_2017	Magnitude of LandTrendr observed shock to NDVI	Sudden positive or negative shock (drought, land use change)
LS_distr_dur_2012_2017	Duration of observed LandTrendr shock to NDVI	Indicator of severity/duration of shock

3.4.2. LandTrendr

LandTrendr (LT) is a broadly used algorithm that detects sudden shifts in an index. For this study we examined NDVI, on a pixel-by-pixel basis. Due to it's lack of importance in the final models, we have moved the description of this data product to the appendix.

3.5. Model results and diagnostics

Table 5 shows the performance metrics of household level models in predicting average monthly income. We show results for the separate machine learning models used, as well as for the set of variables the model can access: (1) satellite variables and survey variables, (2) variables in the LFS and census surveys only, and (3) the satellite derived variables only. We consider two performance metrics: root mean squared error (RMSE), which gives a measure of average error interpreted in units of log household income, and R^2, which measures the coefficient of determination between predicted household income $\hat{y}_h$ and true household income y_h. Model performance metrics at the household level are shown in Table 5. Note that RMSE values indicate a better model performance. An R^2 close to 1 indicates a perfect relationship between the predicted and true household income, and a value of 0 indicates no relationship between predicted and true.

To better characterize our true out-of-sample performance, we separated our dataset into two groups. The first 75% of observations, called the 'training' set, is used to fit all models. The remaining 25% of observations are held out as an independent 'testing' set and provides a much more realistic measure of model performance. Comparing the performance in the training and test sets, we note that cross-validated in-sample performance on the training set overstates model performance. R^2 values indicate we explain between 35% and 60% of the variation in household income when just looking at performance in the training set (using the survey & satellite variables). Out-of-sample performance in the testing set dropped as expected. R^2 values indicate that we can explain 31%–35% of variation in survey and satellite variable models. Much of this is due to a random forest model that appears to be overfit. The performance in the training set for the rest of the models is only slightly higher than that in the test set.

3.6. Model performance with and without satellite variables

Comparing across the set of variables employed at the household level, we see, unsurprisingly, that models with the most variables – survey and satellite – tend to perform best. This is followed by models that use information only available in the survey, which have R^2 values that vary between 0.31 and 0.35, indicating we can reliably explain between 31% and 35% of the variation in household income using survey variables alone. In comparison, models that use only contextual and time series satellite derived variables can explain 10%–14.1% variation in household level income at the household level. Survey variable only models explain around 30%–33% of the variation in the test set. The difference between the survey and survey and satellite variable sets is most pronounced between the

Table 5. Household level model performance, testing and training sets, varying set of Variables used by machine learning models.

	Test (25%, N = 912)		Train (75%, N = 2746, Spatially Cross-Validated at ED Level)				
			RMSE		R2		
Model	RMSE	R2	Mean	SD	Mean	SD	Variable Set
Elastic Net	0.6667	0.317	0.6918	0.032	0.35	0.069	Survey & Satellite
Ridge	0.6622	0.337	0.6522	0.169	0.416	0.094	Survey & Satellite
Extreme Gradient Boosted Trees	0.6533	0.344	0.6592	0.042	0.406	0.071	Survey & Satellite
Random Forest	0.6832	0.281	0.5683	0.04	0.6	0.067	Survey & Satellite
Combined (Average of Models)	0.6505	0.349	0.6286		0.474		Survey & Satellite
Elastic Net	0.6751	0.302	0.6951	0.028	0.337	0.075	Survey
Ridge	0.6763	0.302	0.6945	0.125	0.338	0.059	Survey
Extreme Gradient Boosted Trees	0.6705	0.316	0.6623	0.035	0.399	0.065	Survey
Random Forest	0.6575	0.335	0.5833	0.038	0.573	0.087	Survey
Combined (Average of Models)	0.663	0.325	0.653		0.423		Survey
Elastic Net	0.7651	0.1	0.7938	0.041	0.146	0.071	Satellite
Ridge	0.7607	0.124	0.7495	0.62	0.229	0.069	Satellite
Extreme Gradient Boosted Trees	0.7553	0.135	0.7354	0.062	0.259	0.047	Satellite
Random Forest	0.7598	0.135	0.7126	0.042	0.305	0.062	Satellite
Combined (Average of Models)	0.7473	0.141	0.7356		0.269		Satellite

combined models. For the combined models, the survey and satellite variable set shows an R^2 of 0.349, while for the survey only models we find an R^2 of 0.325. This is a roughly 7.3% improvement in predictive performance.

3.7. Machine learning model comparison

Across machine learning models we see high levels of variability across training set model performance. Meanwhile, actual performance in the test set is more consistent between machine learning models. In the test set R^2 values vary between 0.31 and 0.35, for the preferred models using satellite and survey variables. The best performing individual model is Extreme Gradient Boosted Trees, with an R^2 of 0.344, followed by the Ridge model with an R^2 of 0.337. ElasticNet performs worse with a test-set R^2 of 0.317, and finally random forest performs the worst with an R^2 of 0.281. The highest performing model overall is the combined model which averages all the model predictions, with combined R^2 score of 0.349, and an RMSE value of 0.6505 indicating that our hypothesis of model ensembling does indeed produce the best possible model. Note that there is some evidence that despite the spatially stratified cross-validation, we still see evidence of overfitting. This varies by set of variables used and by modeling approach. Random forest here seems most prone to overfitting. Given that the number of satellite variables greatly outnumbers the survey variables, we should not be surprised that overfitting is possible with including satellite variables. This highlights the importance of using machine learning to regularize and setting up appropriate test sets for model validation.

These plots show estimated average income versus true average income using the combined model averaging estimates between the Ridge, Elastic Net, Random Forest, and Extreme Gradient Boosted Trees models. Note the training sample is the data on which our model has been estimated, and the test data sample is the validation sample, which was not used to directly estimate our model.

3.8. Enumeration district level prediction comparisons

Satisfied with a household level model of income that performs well, we then generate predictions for every household and average these at the ED level to generate average income at this level. Results in the previous analysis led us to believe the best performing model is the combined model, which averages household income predictions across the four machine learning models.

Figure 4 shows predicted versus true plots at the enumeration district level in the training sample of the data. Each point shows the predicted average income for an ED (on the x-axis) against the true average income (on the y-axis). We scale the size of each point by the total population in each ED, and color code them by district. The figure for the testing dataset shows some variation along the 45 degree line, although the fit is altogether fairly strong. Taken together, Table 6 shows the performance metrics at the enumeration district level, comparing the predicted versus true average income levels of the EDs. We see that the R^2, or the coefficient of variation between predicted average ED incomes is highest when using both the satellite and survey variables. In the test set, the most reliable measure of out-of-sample performance, we see an R^2 value of 0.537 using both satellite and survey variables. This falls to 0.497 when we use survey variables only. When we use only satellite variables, we find an R^2 performance of about 0.298, indicating satellite variables alone explain 30% of the variation in ED level average incomes. The RMSE, or root mean squared error, echoes what we see when looking at the R^2 metrics. Again, we see the large decline in performance from the training versus test sets, indicating a clear need to use testing sets as proper validations of performance.

4. Model validation and residuals diagnostics

4.1. Variable importance

Exploring which variables, derived from satellites or surveys, have the most impact on the predictive power of the models is complicated by the fact that there are so many variables, and the variables

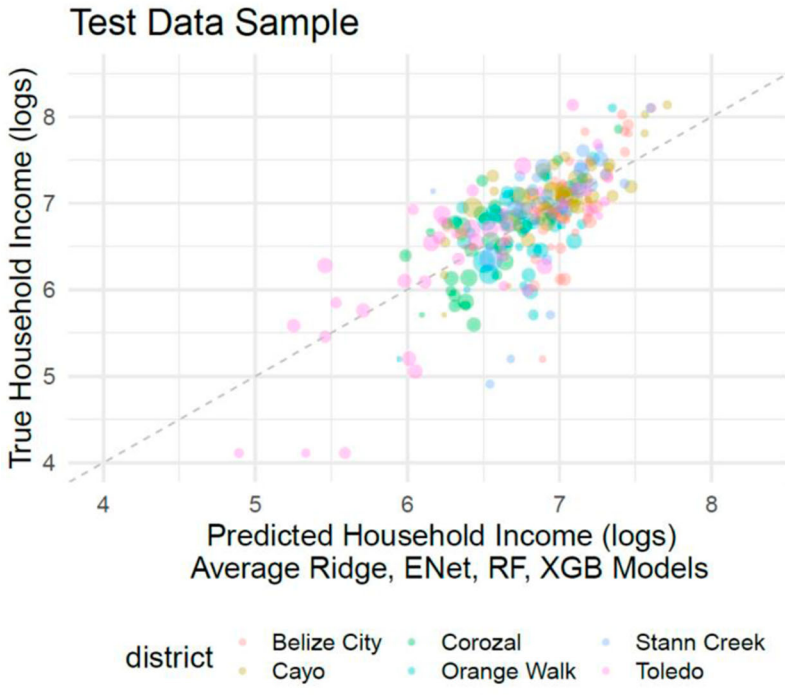


Figure 4. Predicted versus true plot for ED level household income predictions.

Table 6. Enumeration District Level Performance Metrics. Metrics compare true ED average income with predicted average ED income using the 'Combined' model, with variable sets.

R2	RMSE	Data	Variable Set
0.537	0.448	Test	Satellite & Survey
0.298	0.549	Test	Satellite Only
0.497	0.466	Test	Survey Only
0.827	0.229	Train	Satellite & Survey
0.814	0.255	Train	Satellite Only
0.611	0.334	Train	Survey Only

interact non-linearly in our tree-based models (Random Forests and Extreme Gradient Boosted Trees). With simple linear models applied to a more reduced set of variables, we can view estimated coefficients and the resulting t-statistics to determine statistical significance. This is not feasible here given the large set of variables used. One way we can examine which variables improve model performance is by calculating variable importance. This at least provides the relative importance for each variable within a given machine learning model. Variable importance for linear models – Ridge and Elastic Net – is calculated as the absolute value of the t-statistics, which is the coefficient divided by the standard error. For the tree-based models, the variable importance is calculated by averaging all the trees that do not contain a particular variable[5], and comparing mean decrease in final classification purity (or accuracy) against models that do contain these variables. Again, we cannot compare how much each variable improves the models across model types, but we can compare within models, thus we scale the variables importance scores within a model such that 100 is the most important variable, and 50 is half as important as the most important variable.

In the appendix, Figures A2 and A3 presents the variable importance metrics across the four machine learning models, plotting the top 30 most important variables for each. The Ridge,

Random Forests, and Extreme Gradient Boosted Trees select head of household years of education as the most important variable. Elastic Net models also considers this variable important, selecting it third behind number of dependents and number of children. Next most important variables are a set of asset variables – whether a household has high quality cooking fuel, toilet, number of computers, cable access, TV, refrigerator, number of vehicles, and electric washers.

Following these variables, we see several satellite derived variables appear, both generated from the time-series as well as contextual cross-sectional information. For the time-series variables, the longest period below or above mean precipitation tend to be important variables, indicating that areas with long periods of drought or excess rain correlates with changes in income, likely in agricultural areas. Similarly, streaks of NDVI above or below the average level for these areas tend to be predictive of income. Contextual satellite information also appear to be strongly predictive of average incomes. Elastic Net models pick many of these variables – NDVI at 3 pixel scale, SFS at 31 and 71 pixel scale, oriented fast and rotated brief (ORB) at scale 71 pixels, and local binary pattern moments at 3 pixel scale.

Overall, while variables derived from surveys tend to be the strongest predictors of household income, almost all machine learning models improve with the addition of satellite derived variables, and within models outside of the top 5 or 10 most important variables, satellite features tend to be strongly predictive of household income.

4.2. Residuals diagnostics

One concern with small area estimates is that our model may be biased, not on average, but for particular sub-populations whose outcomes we would like to measure with high precision. For instance, we need to ensure we are not producing biased estimates of incomes for the poorest populations. Therefore, residuals diagnostics by subgroups is a crucial component of any small area estimation model.

Figure 5A presents, for the survey only models, the average residuals (true household income minus predicted income) and standard errors by household income decile in the test sample. We use the predictions from the combined model, which calculates the predicted household income as the mean of the four machine learning models. For each income decile, we calculate the average (bar graph length) and the estimated standard error (black bar). Note that the lowest errors are seen for households in the middle of the income distribution, from log incomes of 6.38–7.22. For households within this range, errors are roughly symmetric around zero, and small in magnitude. This indicates we can assume, for households within this income range, our models of income are accurate and unbiased.

As we move to the two highest and lowest two income deciles, we see the average residuals grow. For the richest households, residuals are positive, indicating we underpredict income for these households. For the poorest households, residuals are negative, indicating we overpredict incomes for these households. In general, our models of poverty tend to understate the true variance of household incomes. However, in comparing the top panel A (survey variables only) with the bottom panel B (satellite and survey variables) we see that the residuals for the poorest and richest households are smaller for the satellite and survey models. In particular, residuals for the poorest decile decline from around −0.7 to −0.6. We see similar improvement in other deciles when comparing the satellite and survey to the survey models alone, indicating satellite features help predict rich households as well. Taken together, the satellite features help recover critical characteristics of the most important income deciles.

These plots show average error by actual income quintile. Black bars show estimated standard errors specific to each quintile.

4.3. Maps

Once we are satisfied with an appropriate model, we use the model calibrated from the LFS to predict into the Census. The predictions are averaged by ED to generate average monthly

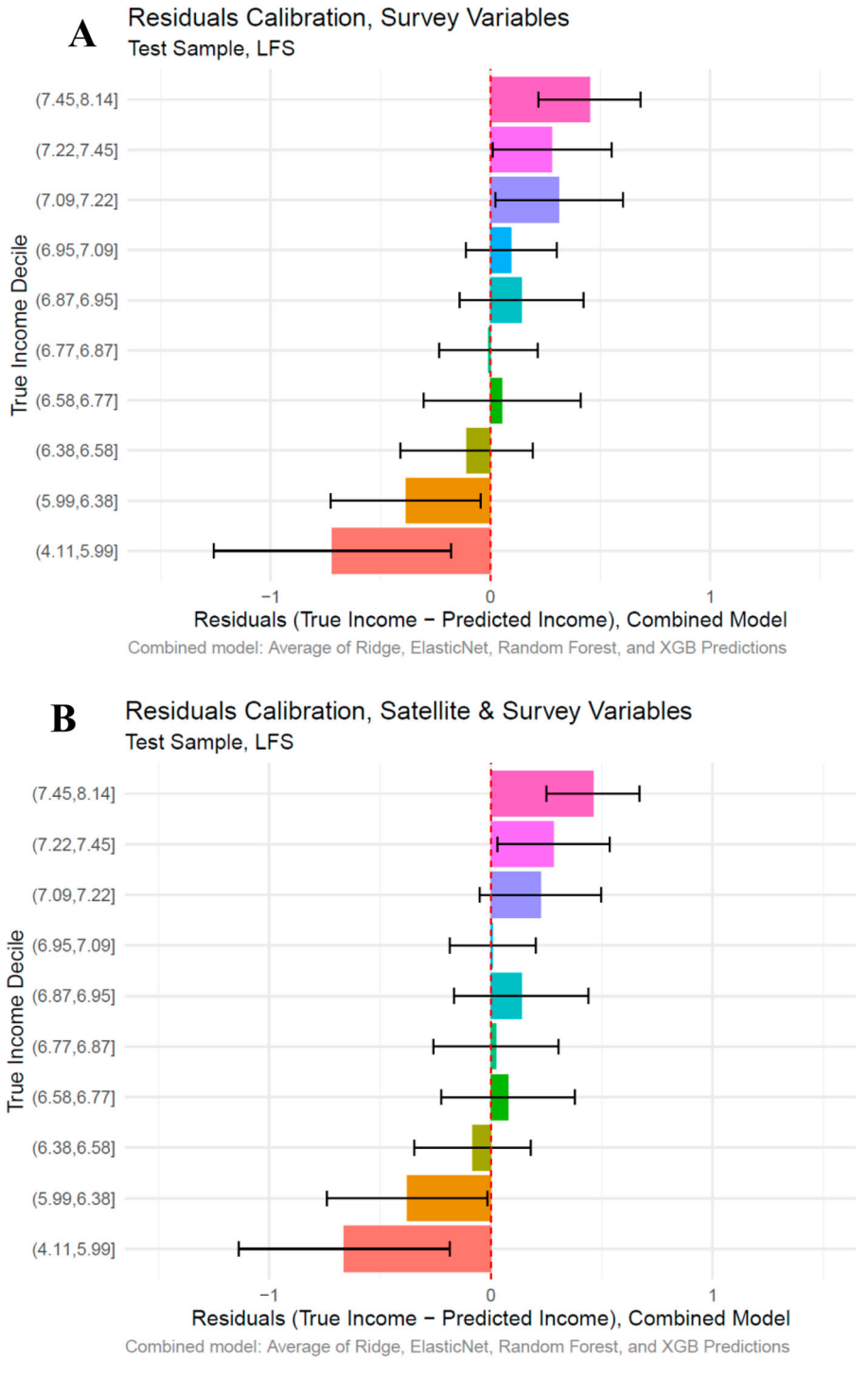


Figure 5. A-B Residuals calibration, survey only models and satellite & survey models.

income, maps of which are shown in the Appendix section of this text. Overall there is a large amount of agreement across models. EDs that appear poor according to the Random Forest model, also appear poor, for the most part, for the Extreme Gradient Boosted Trees model.

Where they differ most significantly is in variance. The linear models – Ridge and ElasticNet – appear to be more constrained in their predictions compared to the XGBoost and Random Forest models. This may have to do with the nonlinear properties of the latter, tree based models, are better able to recover the true income distribution.

To calculate poverty scores, we need to choose a poverty line and calculate headcount poverty rates, or FGT0 counts (Foster et al., 1984). Because of data release restrictions, we calculate poverty using a poverty lines at the 20th percentile of national income. The modeling calibration section showed the best model performance in the test set was the combined model, which averaged all household predictions together. However, we do not have definitive evidence that we have identified the correct machine learning model. This is often referred to as 'model uncertainty' (Chatfield, 1995).

We calculate what we define as the ensemble poverty metric, as shown in the equation below. Suppose we want to know the poverty status of a household h in cluster c. For each household, we have calculated predicted income or consumption for M possible models. The ensemble poverty metric averages across these M models. This better incorporates model uncertainty and will capture the distinction when one model agrees that a household is poor, while three other models do not.

$$\text{Ensemble\_pov}_{h,c} = \frac{1}{M}\sum_{m=1}^{M} 1(\widehat{y_{h,c}^{m}} < \text{PovLine}) \tag{4}$$

The map showing poverty rate using these poverty lines at the 20th percentiles of national income is presented in figure form in the appendix. We see from these figures that the poorest districts are Corozal, in the north, and Toledo, in the south. Viewing the last poverty map that was completed in 2009, four districts were classified as having high poverty rates – Corozal, Orange walk, Stann Creek and Toledo. Viewing our analysis in light of the previous map, it appears there has been significant reduction in poverty for the districts of Orange Walk and Stann Creek. Our new poverty map has enumeration district as their resolution, and the previous maps only provide district level poverty disaggregation, thus it is difficult to make a direct comparison

Given the higher resolution of this poverty map – at the enumeration district rather than district level – other interesting patterns emerge as well. Within a poorer region such as Stann Creek we see there is substantial heterogeneity in the relative poverty rate. The city of Dangrige appears significantly less poor than the surrounding areas. Whether this indicates improvement from 2009 we cannot say, as the previous poverty map did not produce enumeration district poverty levels. The city of Punta Gorda in Toledo, itself a relatively poorer district, appears to have lower poverty and is surrounded by higher poverty EDs. Interestingly the city of Corozul in Corozal still appears as poor as surrounding areas, a pattern that is different from the previous districts discussed.

Given the fact that satellite features have been shown in the previous section to improve the modeling of rural (poorer) households, it's likely that the inclusion of satellite variables allows for the increased accuracy of rural households. This allows us to see details in the poverty rates of households surrounding cities and not restricting to the cities themselves.

5. Discussion and conclusion

What can the proto-typical department of statistics learn from this example in Belize? We have shown that open-source satellite features have the potential to improve poverty estimation, and that they can be incorporated into a machine learning prediction framework with relative ease. We document a 7% improvement in average coefficient of variation when these models incorporate open-source satellite features. However, these averages do not tell the whole story. The model improves the accuracy of income estimates for the poorest households when we add these satellite variables, indicating they hold much potential for identifying the poorest of the poor. This ability to better differentiate the

poorest communities, at a high degree of spatial resolution, is critical to the meaningful targeting of poverty interventions.

The improvements are all the more relevant given that our machine learning models do not come with the same technical debt as those from models built using proprietary data (Sculley et al., 2015). Just as governments are concerned about their level of sovereign financial debt, they should be concerned with level of technical debt within their operational IT systems. Open data for development is proposition that countries should have a bias towards reducing reliance on closed data pipelines they themselves do not control, and further that they release official statistics on their development process with expediency given constraints on quality and as technical limitations allow.

How can international organizations support these developments? First, there is an opportunity for a multilateral development bank or other international organization to openly provide processed spatial data to be included for modeling purposes. Much of the specialized satellite imagery processing is outside the knowledge space of traditional statistical agencies. International organizations can provide these public goods for several countries at a time. One may imagine a time when an API to these features are readily available for several countries. Secondly, as researchers we can support these initiatives by openly posting and sharing our code on Github or other repositories as we have done with our code.[6] Thirdly, we can support statistical capacity initiatives that promote training in the use of openly available data.

We have seen tremendous advances in machine learning that allow countries to produce development metrics more cheaply and with a higher frequency than traditional methods. The methods described here provide a blueprint for leveraging these advances which will hopefully lead to better policy and poverty targeting.

Notes

1. As this article was going to press, we became aware of the paper by Yeh et al. (2020) which also uses publicly available satellite imagery for poverty prediction.
2. See http://iresearch.worldbank.org/PovMap/PovMap2/PovMap2Manual.pdf.
3. For more detail on the methods we refer the reader to the book James et al. (2013) available at: http://faculty.marshall.usc.edu/gareth-james/ISL/.
4. One may also repeat this procedure multiple times, a process known as repeated cross validation (Kim, 2009). While these analyses reflect performing this spatial stratification procedure once, we have repeated the procedure and found hyper-parameters and performance metrics to be stable. Results are available upon request.
5. Recall that in these models variables are randomly selected at each node and therefore some trees will not contain particular variables.
6. Which we have done on our pages https://github.com/mmann1123 and https://github.com/jonhersh. We regret that due to privacy concerns we cannot share the underlying poverty data.
7. http://caret.r-forge.r-project.org/.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Inter-American Development Bank: [Grant Number PEC Contract No. 0002].

Notes on contributors

Jonathan Hersh is an Assistant Professor of economics and management science at the Argyros School of Business at Chapman University. His research interests include the economics of information systems, digital transformation of businesses, technology in developing economies, and the application of predictive models on business and management processes. His work has been featured in *MIS Quarterly,* and *IZA World of Labor*. He frequently consults with

businesses and international organizations on how to build predictive models to augment data-driven decision making. Previously he was a lecturer at MIT and Wellesley College. He is a graduate of the University of Chicago, University of Pennsylvania and Boston University.

Ryan Engstrom is an Associate Professor in the Department of Geography at George Washington University (GWU) where he is the Director of the Center for Urban and Environmental Research (CUER) and the Spatial Analysis Lab (SAL). His research interests are in using geospatial techniques including remote sensing, spatial statistics, and Geographic Information Systems (GIS) to understand spatial variations in a wide array of issues. He has worked on projects focusing on poverty, climate change, health, and population in a range of geographic areas including the Arctic, Africa, Asia, and Washington, DC. He has been funded by and collaborated with a wide range of institutions including NASA, NSF, NIH, World Bank, Children's National Medical Center, USAID, Inter-American Development Bank (IADB), and the Ford Foundation. He earned his Ph.D. in Geography from the joint program between San Diego State University and University of California, Santa Barbara in 2005.

Michael Mann is an Assistant Professor of Geography at The George Washington University. Here he teaches classes on GIS, Python and R computing languages, and spatial modeling. His research has focused on the application of spatial data, and econometric and machine learning techniques to forecasting of human/natural systems interactions. This has included modeling the economic determinants of deforestation in Brazil, forecasting housing development in California and economic losses due to wildfire, and agricultural modeling in Ethiopia and India. He also has interests in remote sensing, high performance computing, data visualization, and web mapping. Michael received his BA in Economics from College of Wooster in Ohio. He then received his PhD in Geography and an MA in Environmental and Policy Modeling at Boston University. This was followed by postdoctoral research at UC Berkeley in a wildfire ecology lab.

ORCID

Jonathan Hersh http://orcid.org/0000-0001-6786-5162

References

Afzal, M., Hersh, J., & Newhouse, D. (2015). *Building a better model: Variable selection to predict poverty in Pakistan and Sri Lanka*. Mimeo, World Bank.

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal*, *99*(394), 116–131. https://doi.org/10.2307/2234208

Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A., & Swartz, T. (2017). Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico. ArXiv Preprint ArXiv:1711.06323.

Bahn, V., & McGill, B. J. (2013). Testing the predictive performance of distribution models. *Oikos*, *122*(3), 321–331. https://doi.org/10.1111/j.1600-0706.2012.00299.x

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350* (6264), 1073–1076. https://doi.org/10.1126/science.aac4420

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

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *158*(3), 419–444. https://doi.org/10.2307/2983440

Cohen, W. B., Yang, Z., Healey, S. P., Kennedy, R. E., & Gorelick, N. (2018). A LandTrendr multispectral ensemble for forest disturbance detection. *Remote Sensing of Environment*, *205*, 131–140. https://doi.org/10.1016/j.rse.2017.11.015

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009, June 20–June 25). *Imagenet: A large-scale hierarchical image database*. 2009 IEEE conference on computer vision and pattern recognition.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro–level estimation of poverty and inequality. *Econometrica*, *71*(1), 355–364. https://doi.org/10.1111/1468-0262.00399

Engstrom, R., Hersh, J., & Newhouse, D. (2017). Poverty from space: Using high-resolution satellite imagery for estimating economic well-being.

Engstrom, R., Pavelsku, D., Tomomi, T., & Wambile, A. (n.d.). Mapping Poverty and Slums Using Multiple Methodologies in Accra, Ghana. In Joint Urban Remote Sensing Conference 2019 (pp. 1–4).

Foster, J., Greer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, *52*, 761–766. https://doi.org/10.2307/1913475

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*, 1189–1232. https://doi.org/10.1214/aos/1013203451

Funk, C. C., Peterson, P. J., Landsfeld, M. F., Pedreros, D. H., Verdin, J. P., Rowland, J. D., Romero, B. E., Husak, G. J., Michaelsen, J. C., & Verdin, A. P. (2014). A quasi-global precipitation time series for drought monitoring. *US Geological Survey Data Series*, *832*(4), 1–12. https://doi.org/10.3133/ds832.

Groger, A., Hersh, J., Matranga, A., Mueller, H., & Serrat, J. (2020). Analyzing conflict from space: Identification of physical destruction during the Syrian civil war with artificial intelligence.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12* (1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Huang, X., Zhang, L., & Li, P. (2007). Classification and extraction of spatial features in urban areas using high-resolution Multispectral imagery. *IEEE Geoscience and Remote Sensing Letters*, *4*(2), 260–264. https://doi.org/10.1109/LGRS.2006.890540

Jacques, D. C. (2018). Mobile phone metadata for development. arXiv preprint arXiv:1806.03086.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). springer.

Jean, N., Burke, M., Xie, M., Matthew Davis, W., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Kennedy, R. E., Yang, Z., & Cohen, W. B. (2010). Detecting trends in forest disturbance and recovery using yearly landsat time series: 1. LandTrendr—temporal segmentation algorithms. *Remote Sensing of Environment*, *114*(12), 2897–2910. https://doi.org/10.1016/j.rse.2010.07.008

Kilic, T., Serajuddin, U., Uematsu, H., & Yoshida, N. (2017). *Costing household surveys for monitoring progress toward ending extreme poverty and boosting shared prosperity*. The World Bank.

Kim, J. H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735–3745. https://doi.org/10.1016/j.csda.2009.04.009

Kirscht, M. (1998). Detection, velocity estimation and imaging of moving targets with single-channel SAR. In *EUSAR'98-European Conference on Synthetic Aperture Radar* (pp. 587–590), Friedrichshafen, Germany.

Kirscht, M., & Rinke, C. (1998, November 17–19). *3D Reconstruction of buildings and vegetation from Synthetic Aperture Radar (SAR) images*. In MVA, IAPR Workshop on Machine Vision Application, Makuhari, Chiba Japan, 228–231.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Maail, A. G. (2017). Linking open data and the fight against corruption in Indonesia. *The Indonesian Journal of Development Planning*, *I*(3-Dec 2017). https://doi.org/10.36574/jpp.v1i3.23.

Mann, M. L., & Warner, J. M. (2017). Ethiopian wheat yield and yield gap estimation: A spatially explicit small area integrated data approach. *Field Crops Research*, *201*, 60–74. https://doi.org/10.1016/j.fcr.2016.10.014

Mann, M. L., Warner, J. M., & Malik, A. S. (2019). Predicting high-magnitude, low-frequency crop losses using machine learning: An application to cereal crops in Ethiopia. *Climatic Change*, *154*(1–2), 211–227. https://doi.org/10.1007/s10584-019-02432-7

McNairn, H., Shang, J., Jiao, X., & Champagne, C. (2009). The Contribution of ALOS PALSAR multipolarization and polarimetric data to crop classification. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(12), 3981–3992. https://doi.org/10.1109/TGRS.2009.2026052

Mehrotra, R., Namuduri, K. R., & Ranganathan, N. (1992). Gabor filter-based edge detection. *Pattern Recognition*, *25*(12), 1479–1494. https://doi.org/10.1016/0031-3203(92)90121-X

Myint, S. W., Mesev, V., & Lam, N. (2006). Urban textural analysis from remote sensor data: Lacunarity measurements based on the differential box counting method. *Geographical Analysis*, *38*(4), 371–390. https://doi.org/10.1111/j.1538-4632.2006.00691.x

National Human Development Advisory Committee, Ministry of Economic Development, Commerce and Industry, and Consumer Protection. (2010). 2009 Country poverty assessment. http://www.ncabz.org/wp-content/uploads/2015/05/Country-Povert-Assessment.pdf

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987. https://doi.org/10.1109/TPAMI.2002.1017623

Pape, U. J., & Mistiaen, J. A. (2018). Household expenditure and poverty measures in 60 minutes: A new approach with results from Mogadishu. World Bank Policy Research Working Paper, (8430).

Pesaresi, M., Corbane, C., Julea, A., Florczyk, A. J., Syrris, V., & Soille, P. (2016). Assessment of the added-value of Sentinel-2 for detecting built-up areas. *Remote Sensing*, *8*(4), 299. https://doi.org/10.3390/rs8040299

Pesaresi, M., Gerhardinger, A., & Kayitakire, F. (2008). A robust built-up area presence index by Anisotropic Rotation-Invariant Textural measure. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *1*(3), 180–192. https://doi.org/10.1109/JSTARS.2008.2002869

Pokhriyal, N., & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, *114*(46), E9783–E9792. https://doi.org/10.1073/pnas.1700319114

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. R. (2011, Nov 6–13). 2011 International conference on computer vision.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 2503–2511).

Serajuddin, U., Uematsu, H., Wieser, C., Yoshida, N., & Dabalen, A. (2015). *Data deprivation: Another deprivation to end*. The World Bank.

Shan, J., Hussain, E., Kim, K., & Biehl, L. (2010). Flood mapping with satellite images and its web service. *Photogrammetric Engineering & Remote Sensing*, *76*(2), 102–104.

Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., & Lucas, R. (2014). New global Forest/Non-Forest maps from ALOS PALSAR data (2007–2010). *Remote Sensing of Environment*, *155*, 13–31. https://doi.org/10.1016/j.rse.2014.04.014

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T.J., & Blumenstock, J. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, *14*(127).

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment*, *8*(2), 127–150.

Ünsalan, C., & Boyer, K. L. (2005). A System to detect Houses and Residential Street Networks in Multispectral satellite images. *Computer Vision and Image Understanding*, *98*(3), 423–461. https://doi.org/10.1016/j.cviu.2004.10.006

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J. P., Camps-Valls, G., & Moreno, J. (2012). Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sensing of Environment*, *118*, 127–139. https://doi.org/10.1016/j.rse.2011.11.002

World Bank Group. (2017). *World Bank support for open data 2012-2017*. World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/28616 License: CC BY 3.0 IGO

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-16185-w

Yoshida, N., Munoz, R., Skinner, A., Lee, C. K.-E., Brataj, M., Durbin, S. W., & Sharma, D. (2015). *Survey of well-being via instant and frequent tracking (SWIFT) data collection guidelines (English)*. World Bank Group. http://documents.worldbank.org/curated/en/591711545170814297/Survey-of-Well-Being-via-Instant-and-Frequent-Tracking-SWIFT-Data-Collection-Guidelines

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix

Contextual features

In this analysis, we use scales of 3, 5, 7, which are squares of 3 pixels by 3, 5 pixels by 5 pixels, to 7 pixels by 7 pixels for the majority of features. This constitutes looking at an area of 30, 50, and 70 m for the 'neighborhoods' which will constitute the windows of analysis for our contextual features. For the features ORB, SFS, Fourier and LSR the scale was increased by a factor of 10 because these features need more area to properly capture the variation in the landscape.

Each of the contextual features may have several different outputs depending upon the statistical properties of the features as those features are calculated. For Pantex and Lacunarity, the actual values themselves are outputted. For NDVI, Mean, and Fourier, just the mean and variance are outputted. For HOG, LPBPM, and ORB, we output the Mean, Maximum, Vvariance, Kurtosis, and Skewness for these measures. LSR outputs the line contrast, line length, and line mean. SFS outputs maximum line length, minimum line length, mean, weight mean, standard deviation, and maximum ration of orthogonal angles. Finally, Gabor outputs mean and variance for each of the filters, which in this study we used 14.

In total, this produces 46 total outputs for all of the features, and, because each feature is run at 3 scales, in sum our method produces a total sum of 144 outputs from the contextual features. The eventual geographic area to which we link these satellite features is the Enumeration District (ED), thus for each ED area we summarize the features using the mean, standard deviation, and the sum for each. Together this produces 432 contextual feature values, which summarize various contextual aspects of satellite imagery for each ED.

As is displayed in Figure 2, the spatial and spectral patterns of the urban area visible within the imagery is well captured by contextual features.

These features are primitive versions of the features constructed using machine learning techniques such as Convolutional Neural Networks (Jean et al., 2016). Both approaches summarize images by comparing pixels with their neighbors. The main difference is that the Convolutional Neural Networks require survey data on welfare to determine which features to calculate. In other words, the computer selects parameters for layers of filters., which Wwhen applied to the imagery, these parameters construct textures that are optimized to distinguish between low and higher welfare areas. In order for the computer to select the best parameters for these layers, the general method is to use millions of data points when training the algorithms. Because of the limited training data for poverty surveys, a method known as 'transfer learning' is used to hot-start the intermediate layers of the convolutional neural network which defines the filters (Babenko et al., 2017; Jean et al., 2016), often using intermediate filters that have been trained against large corpuses of images

such as ImageNet (Deng et al., 2009). In practice, this assigns filters intended for the purpose of recognizing features in traditional photography to satellite images.

In contrast, the contextual features used in this analysis are constructed using pre-determined algorithms. Therefore, they are independent of the survey data. To be clear, both methods must use external information to inform the choice of filters which summarize the imagery. Both methods use pre-defined filters given the paucity of survey data, only ours are designed for summarizing satellite images and not photographs of dogs and cats.

Landtrendr

LandTrendr (LT) is a broadly used algorithm that detects sudden shifts in an index, in this case NDVI, on a pixel-by-pixel basis. Effectively, LT fits local regressions, usingthe uses a series of metrics to detect sudden shifts in slope, or intercept on a year-by-year basis. As such, LTR is effective at identifying land cover change, for instance conversion of forest to agriculture, or agriculture to urban settlement. Because LT only needs one clear observation per year, it is effective for use with high resolution data like Sentinel or Landsat or another higher resolution satellite. LT has a series of underlying assumptions and parameters, which we will not cover here for the sake of clarity. Detailed information on the algorithm is available (Cohen et al., 2018; Kennedy et al., 2010). For a visual example of LandTrendr's output see Figure A1 below:

Model estimation

Across all four machine learning prediction methods we use spatially stratified (at the district level) leave-one-out cross-validation to tune necessary parameters in the model. The shrinkage parameters for models 1 and 2 are selected via spatially stratified cross-validation. For model 3 we utilize 100 regression trees. We cross-validate over two parameters. First parameter is the number of variables sampled at each split ('mtry') using a grid from 1 to 30, in steps of 5. The second parameter is the minimum number of observations at the end of a leaf at which no more splitting can occur. This helps to control the depth of the tree. For model 4, we cross validate the following parameters: number of trees grown from 50 to 400, maximum tree depth from 1 to 5 in steps of 5, learning rate from 0.3 to 0.4, and variables sampled from 60% to 80%. All models are estimated in R using the package caret.[7]

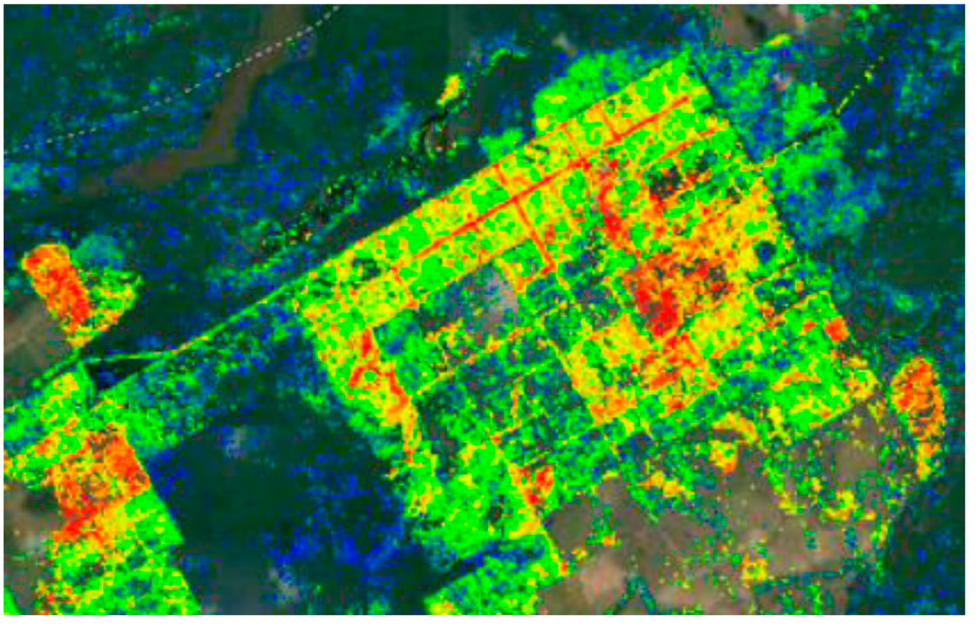


Figure A1. LandTrendr magnitude of disturbance In the image above we can see that LandTrendr's algorithm can detect one-time shocks, such as paving or resurfacing roads, conversion of crop type (or perhaps crop loss). Areas in red indicate a strong negative shock, such as the creation of a road, or reduction in greenness. Areas in light green indicate mild shocks perhaps indicating typical planting, harvest cycles.

Variable importance

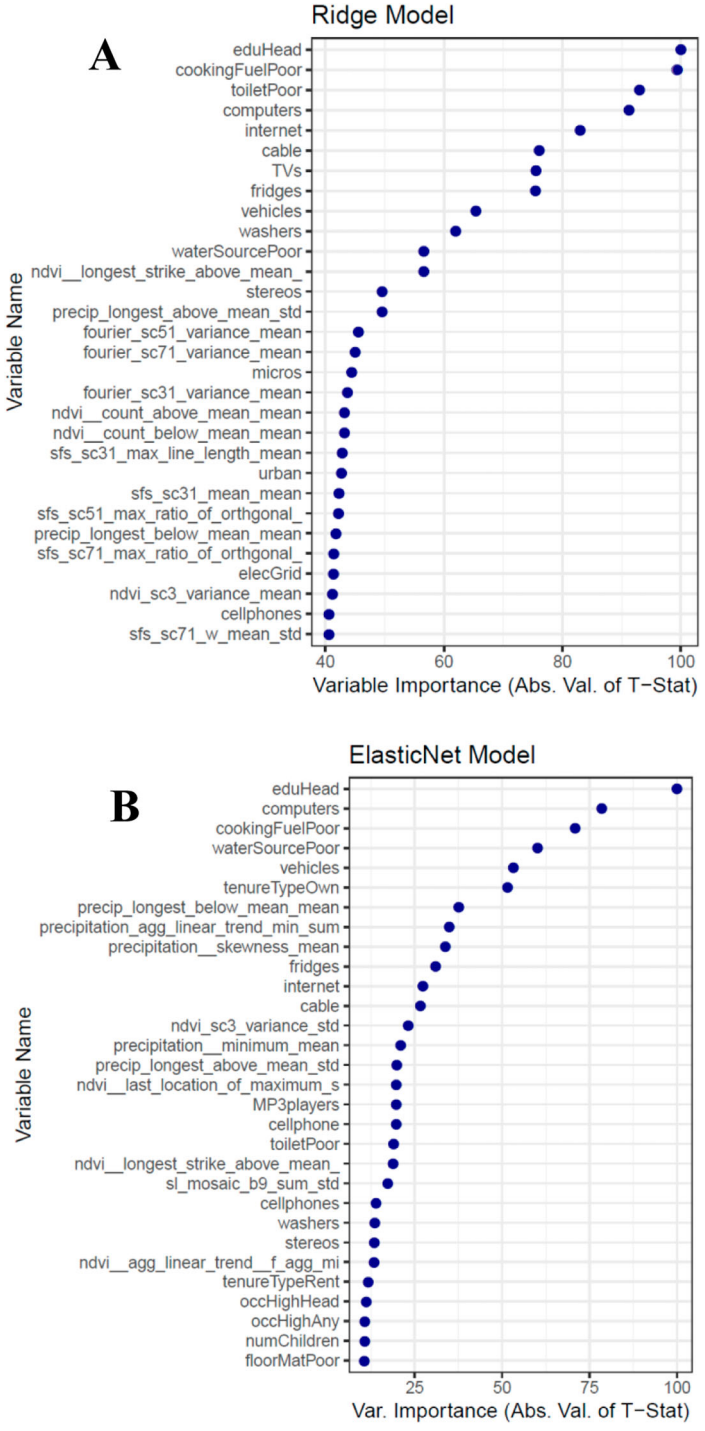


Figure A2. (A–B) Variable importance machine learning models of household income, Ridge and Elastic Net models. These plots show the top 30 most important variables for predicting household income. Each variable's importance metric is scaled by the top variable, which is given an importance metric of 100, and other variables scores are relative to that variable.

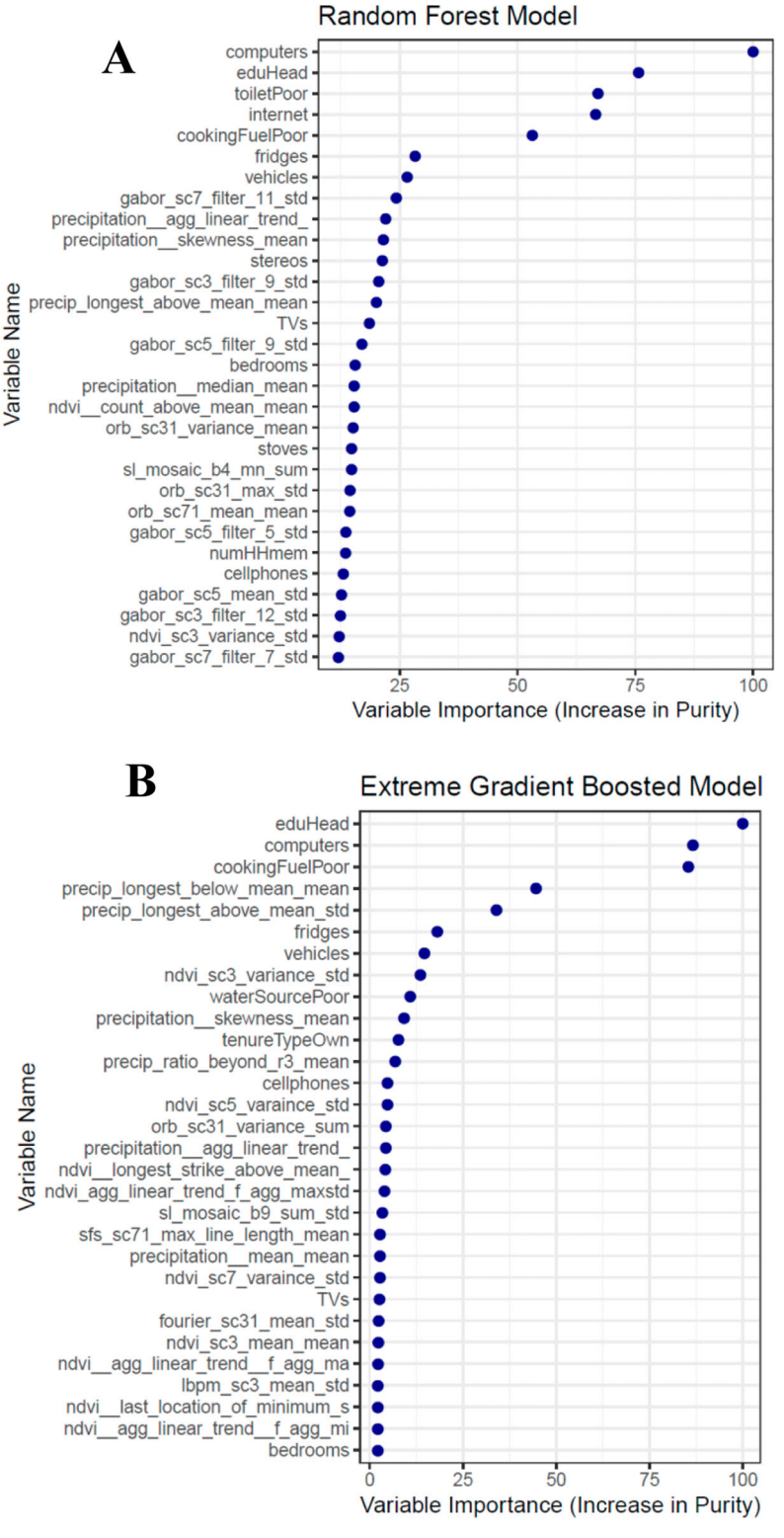
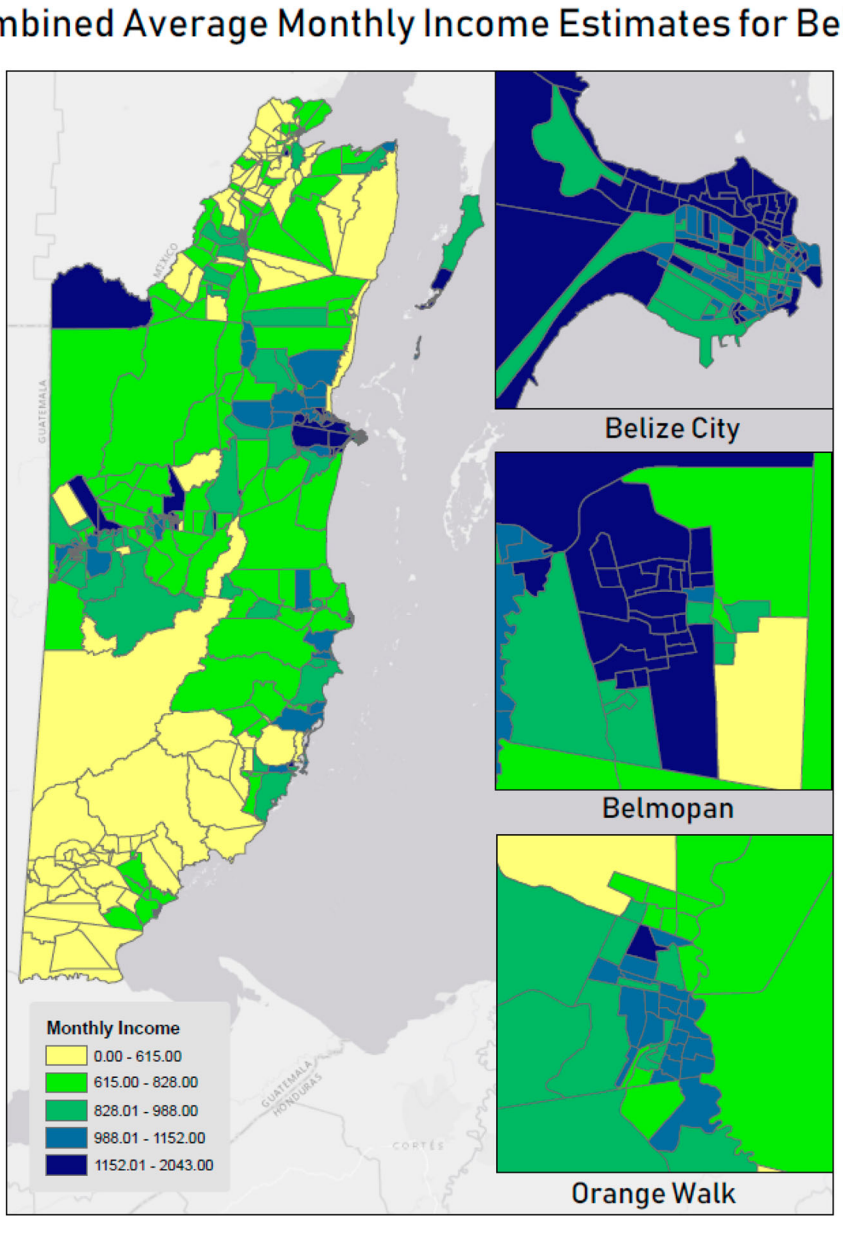
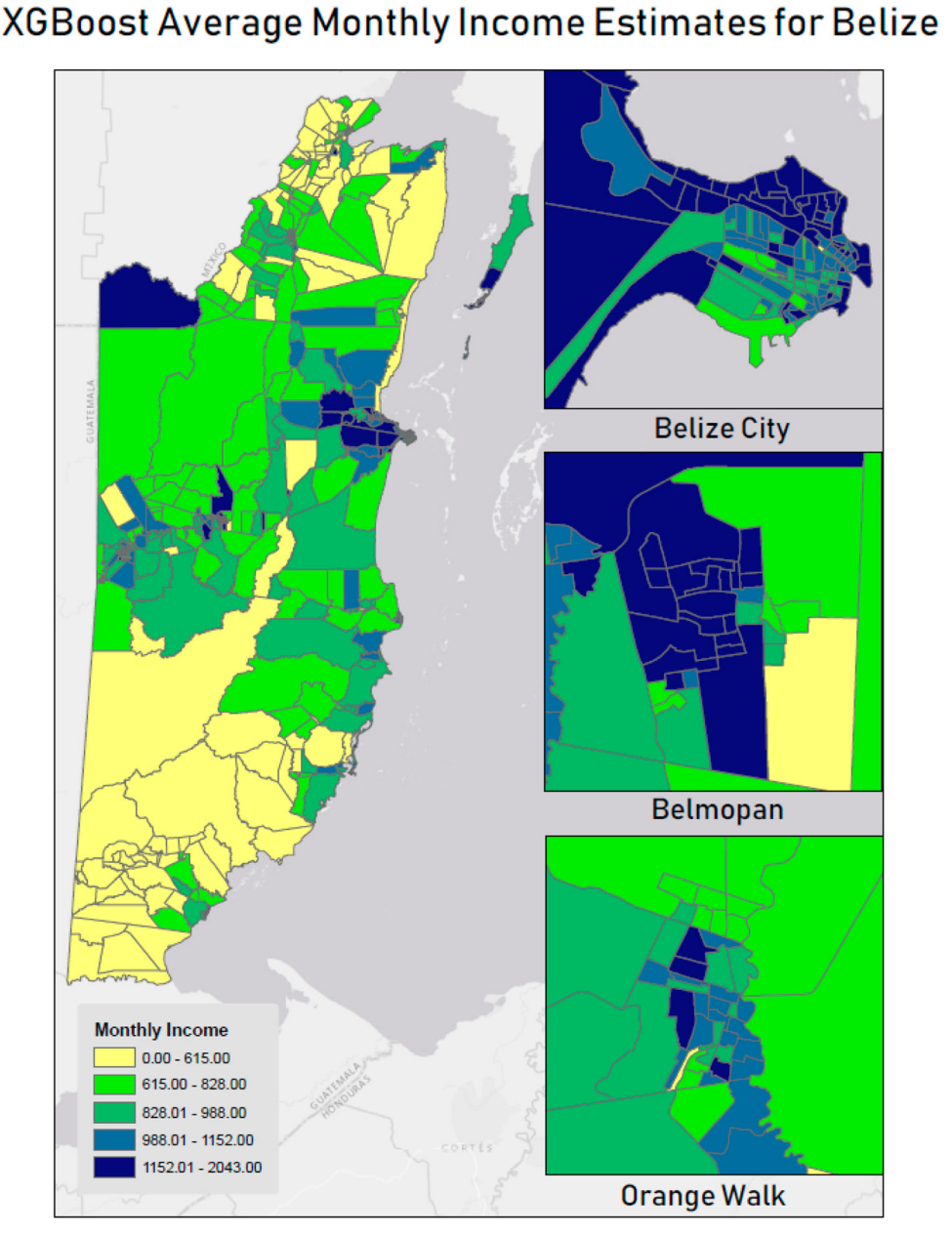
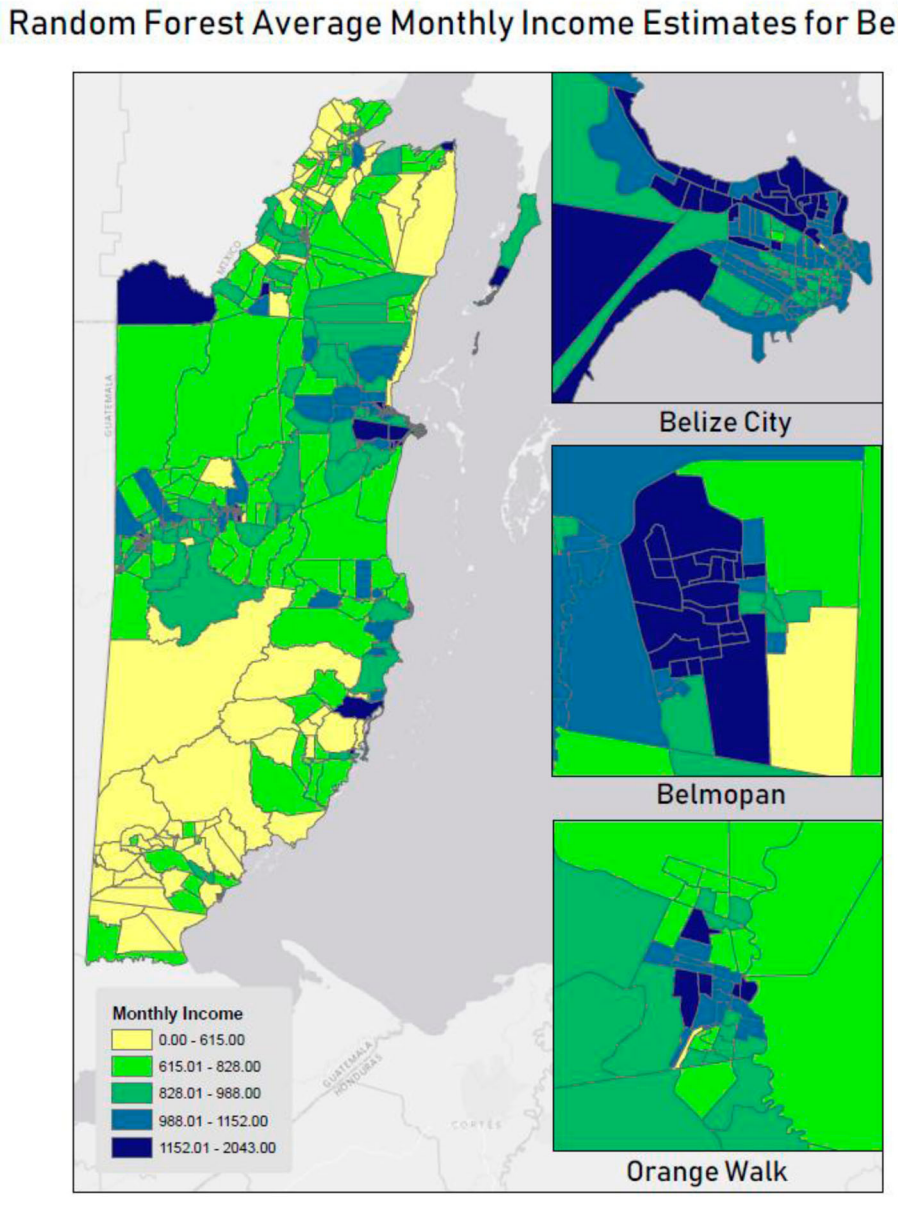


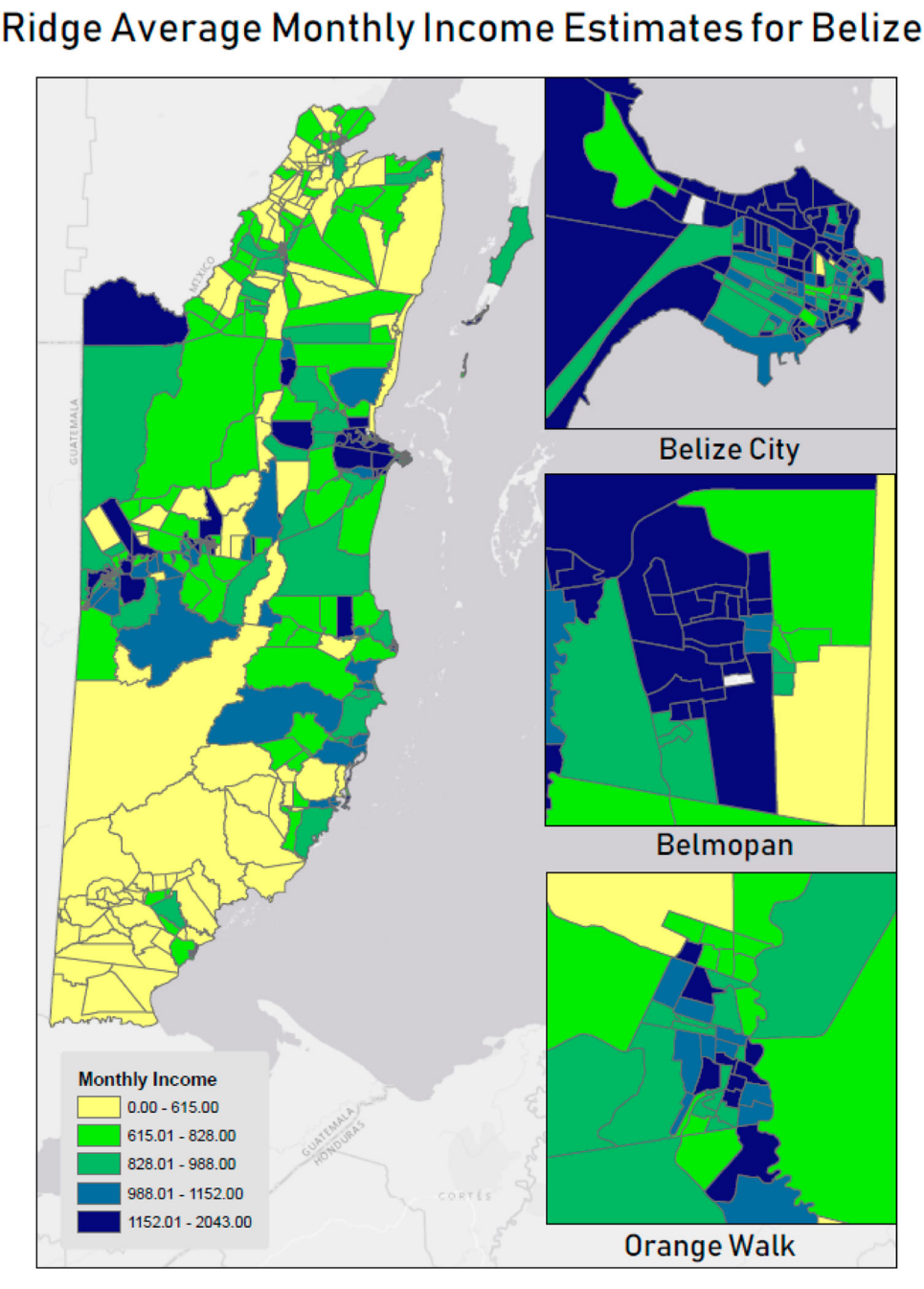
Figure A3. (A–B) Variable importance machine learning models of household income, Random Forest and Extreme Gradient Boosted models. These plots show the top 30 most important variables for predicting household income for Random Forest and XGBoost. Each variable's importance metric is scaled by the top variable, which is given an importance metric of 100, and other variables scores are relative to that variable.

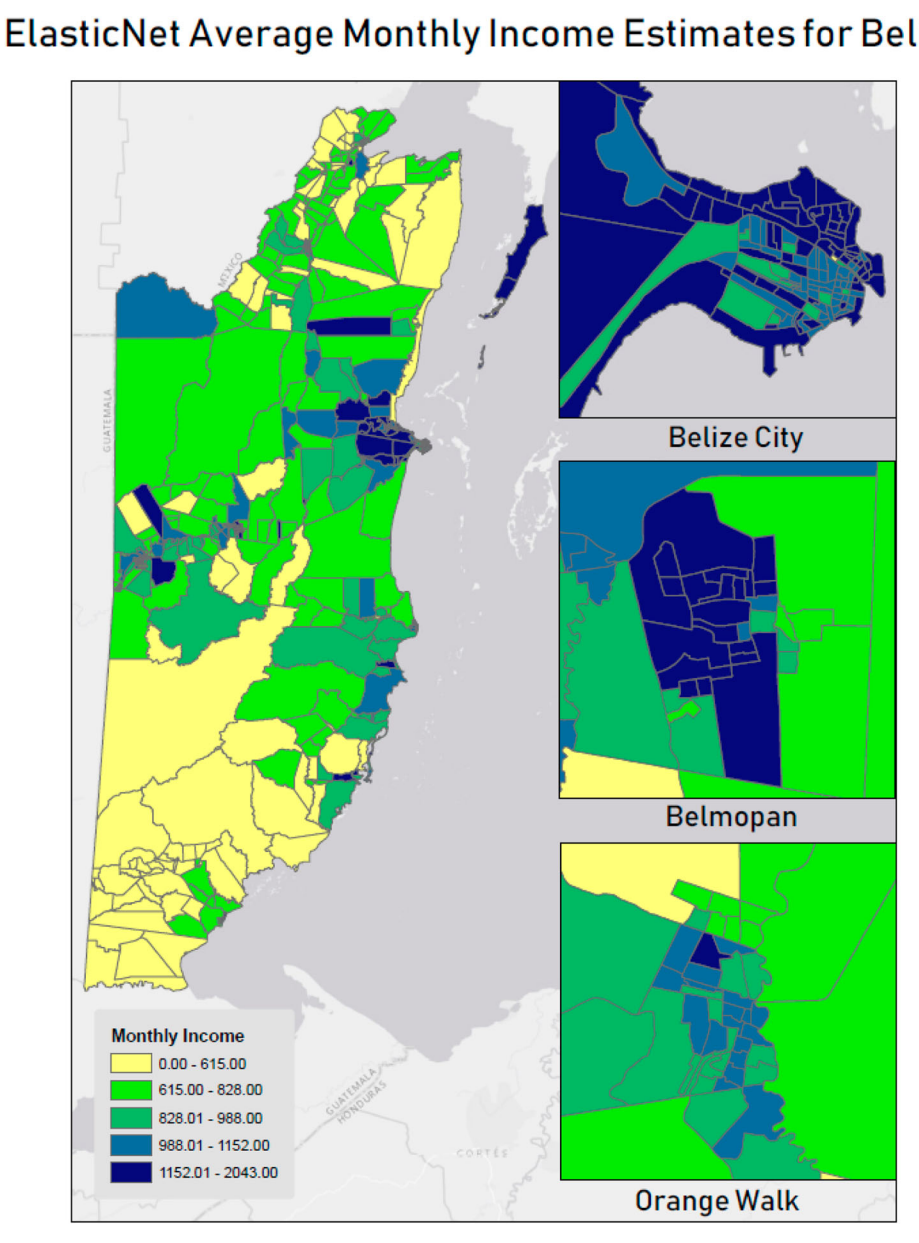
Maps

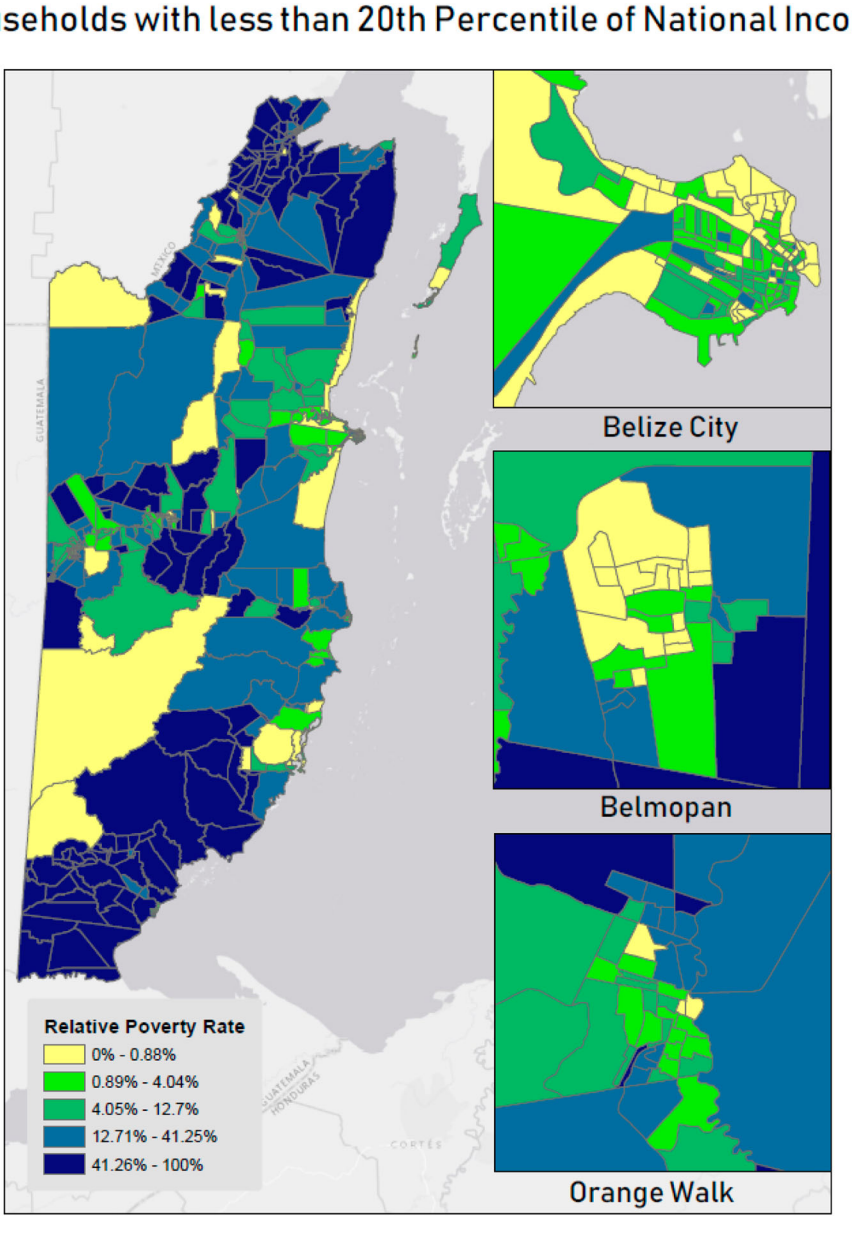

Combined Average Monthly Income Estimates for Belize
Belize City
Belmopan
Orange Walk
Monthly Income
0.00 - 615.00
615.00 - 828.00
828.01 - 988.00
988.01 - 1152.00
1152.01 - 2043.00
MEXICO
GUATEMALA
GUATEMALA
HONDURAS
CORTÉS


XGBoost Average Monthly Income Estimates for Belize
Belize City
Belmopan
Orange Walk
Monthly Income
0.00 - 615.00
615.00 - 828.00
828.01 - 988.00
988.01 - 1152.00
1152.01 - 2043.00
MEXICO
GUATEMALA
GUATEMALA
HONDURAS
CORTÉS


Random Forest Average Monthly Income Estimates for Belize
Belize City
Belmopan
Orange Walk
Monthly Income
0.00 - 615.00
615.01 - 828.00
828.01 - 988.00
988.01 - 1152.00
1152.01 - 2043.00
MEXICO
GUATEMALA
GUATEMALA
HONDURAS
CORTES


Ridge Average Monthly Income Estimates for Belize
MEXICO
GUATEMALA
Belize City
Belmopan
Orange Walk
Monthly Income
0.00 - 615.00
615.01 - 828.00
828.01 - 988.00
988.01 - 1152.00
1152.01 - 2043.00
GUATEMALA
HONDURAS
CORTÉS


ElasticNet Average Monthly Income Estimates for Belize
Belize City
Belmopan
Orange Walk
Monthly Income
0.00 - 615.00
615.00 - 828.00
828.01 - 988.00
988.01 - 1152.00
1152.01 - 2043.00
MEXICO
GUATEMALA
GUATEMALA
HONDURAS
CORTÉS


Households with less than 20th Percentile of National Income
Belize City
Belmopan
Orange Walk
Relative Poverty Rate
0% - 0.88%
0.89% - 4.04%
4.05% - 12.7%
12.71% - 41.25%
41.26% - 100%
MEXICO
GUATEMALA
GUATEMALA
HONDURAS
CORTÉS