Can Human Development be Measured with Satellite Imagery?

Andrew Head
Department of Computer Science
University of California, Berkeley
Berkeley, CA 94720
andrewhead@berkeley.edu

Mélanie Manguin
Department of Industrial Engineering and
Operations Research
University of California, Berkeley
Berkeley, CA 94720
melanie.manguin@berkeley.edu

Nhat Tran
Department of Bioengineering
University of California, San Francisco
San Francisco, CA 94143
ntran16@berkeley.edu

Joshua E. Blumenstock
School of Information
University of California, Berkeley
Berkeley, CA 94720
jblumenstock@berkeley.edu

ABSTRACT

In many developing country environments, it is difficult or impossible to obtain recent, reliable estimates of human development. Nationally representative household surveys, which are the standard instrument for determining development policy and priorities, are typically too expensive to collect with any regularity. Recently, however, researchers have shown the potential for remote sensing technologies to provide a possible solution to this data constraint. In particular, recent work indicates that satellite imagery can be processed with deep neural networks to accurately estimate the sub-regional distribution of wealth in sub-Saharan Africa.

In this paper, we explore the extent to which the same approach—of using convolutional neural networks to process satellite imagery—can be used to measure a broader set of human development indicators, in a broader range of geographic contexts. Our analysis produces three main results: First, we successfully replicate prior work showing that satellite images can accurately infer a wealth-based index of poverty in sub-Saharan Africa. Second, we show that this approach can generalize to predicting poverty in other countries and continents, but that the performance is sensitive to the hyperparameters used to tune the learning algorithm. Finally, we find that this approach does not trivially generalize to predicting other measures of development such as educational attainment, access to drinking water, and a variety of health-related indicators. We discuss in detail whether these findings represent a fundamental limitation of this approach, or could be fixed through more concerted adaptations of the machine learning environment.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; • **Applied computing** → *Economics*;

KEYWORDS

Deep Learning, Convolutional Neural Networks, Transfer Learning, Poverty, Development, Satellite Imagery

ACM Reference format:
Andrew Head, Mélanie Manguin, Nhat Tran, and Joshua E. Blumenstock. 2017. Can Human Development be Measured with Satellite Imagery?. In *Proceedings of ICTD '17, Lahore, Pakistan, November 16–19, 2017,* 11 pages.
https://doi.org/10.1145/3136560.3136576

1 INTRODUCTION AND RELATED WORK

At the United Nations Summit in September 2016, the world committed to the 2030 Agenda for Sustainable Development. The agenda lists the 17 Sustainable Development Goals (SDGs) that balance three key aspects of sustainable development: economic, social, and environmental [20]. The 2030 agenda particularly stresses the need to track more than just country-level Gross Domestic Product (GDP). This requires more timely, reliable, and appropriate ways of collecting and interpreting information on a broad range of human development outcomes.

Measuring human development has long been a focus of international development research and policy [5, 10, 11, 22]. Timely and accurate data can assist government actors in optimally targeting policies and efficiently allocating resources [4]. In the context of the SDGs, it can also provide a useful benchmark for progress, and assist in evaluation of policies. Unfortunately, reliable data is typically very expensive to collect, and thus a major obstacle to effective policy design has been the lack of timely and reliable socioeconomic data [16].

In the past several years, recent developments in machine learning and geospatial analysis have enabled novel data-intensive approaches to the measurement of poverty [7]. Early work relied on satellite "night-lights" data, and showed that regions emitting high levels of artificial light tended to have higher economic output [9, 14]. For instance, Mellander et al. [18] show that such data correlates closely with wage income in Sweden (R^2=0.70), and Noor et

ICTD '17, November 16–19, 2017, Lahore, Pakistan

ACM ISBN 978-1-4503-5277-2/17/11…$15.00
https://doi.org/10.1145/3136560.3136576

al. [21] show that nighttime luminosity correlates with asset-based measures of wealth in 37 African countries. However, this approach generally under-performs in low income regions, because the stable light level often cannot be distinguished from the noise in the data [9]. For example, Jean et al. [15] show that night-lights alone do a poor job of differentiating between poor and ultra-poor regions in sub-Saharan Africa.

The limitations of night-lights data, particularly in poor rural areas, inspired recent papers that use *daytime* satellite imagery [15, 26], raster and vector datasets [23], and mobile phone data [6] to measure poverty in developing countries. One particularly effective approach was recently developed by Jean et al. [15], who use a two-step transfer learning framework to estimate sub-regional levels of asset- and consumption-based poverty in five countries in sub-Saharan Africa. They show that the transfer learning approach outperforms the predictions of night-lights data, improving R^2 by at least 0.10 in more than 70% of independent trials. These results have captured the imagination of many in the development community, as the method utilizes only publicly available data and open source software [19].

The focus of the current paper is to "stress test" the generalizability of this general approach to measuring human development with satellite imagery, to build a more robust evidence base for those wishing to apply these methods in a broader range of development contexts. Specifically, we wish to better understand whether an analogous approach can be used to predict a broader range of outcomes beyond wealth and expenditures, and in particular, whether satellite imagery can also accurately predict key development indicators such as levels of education, access to clean drinking water, and health-related outcomes. We also seek to test whether the same approach can effectively estimate poverty (and other measures of development) in countries outside of sub-Saharan Africa, for instance in the Caribbean or in South Asia.

Our analysis produces several novel results. We begin by replicating the experiments performed by Jean et al., and show that the transfer learning approach can indeed be used to reconstruct accurate wealth indexes in sub-Saharan Africa. We then show that with only modest adaptations, the same formula can be used to achieve reasonable, though attenuated, performance in predicting wealth outside of Africa. However, we find that performance degrades very quickly when trying to use this method to recover other measures of human development. We also show that the performance of the algorithm is sensitive to the hyperparameters used to train the neural network; we interpret this as a cautionary tale that one should not expect to be able to use these algorithms "out of the box," and that they cannot be applied without careful tuning.

Our primary conclusion is thus that this approach to estimating poverty from satellite imagery does not trivially generalize to other measures of human development in other countries. This is a cautionary example to many who look with great optimism to the potential for remote sensing technologies to solve data constraints in international development. However, whether this represents a fundamental limitation of the approach (for instance, that there is simply not enough information in satellite data to infer levels of eduction) or a shortfall of our current efforts (for instance, that the algorithms must be adapted) is an important topic of discussion we return to in the pages that follow, and a question that we hope can motivate future research in this area.

2 DATA

Our analysis leverages data from three different sources. These datasets are summarized in Table 1 and described in turn below.

2.1 Demographic and Health Surveys (DHS)

We rely on the Demographic and Health Surveys (DHS) as a measure of "ground truth" for development outcomes. These nationally-representative household survey data are collected in 90 countries worldwide. In a typical DHS, tens of thousands of households are surveyed on a wide variety of demographic, social, economic, and health-related outcomes [25]. The approximate location of each household is recorded, and the public data is released with a geographic "cluster" assigned to each household. We downloaded the most recent version of the Standard DHS data for four different countries: Rwanda (2010), Nigeria (2013), Haiti (2012), and Nepal (2011). The first panel of Table 1 provides summary statistics of the DHS data in each of these four countries.

Each DHS contains hundreds of questions. We focus our analysis on the following subset of questions in the DHS, which are intended to capture a broad range of development indicators in the spirit of the Sustainable Development Goals.

Wealth. The DHS survey data provides a continuous-scale wealth index, calculated through principal component analysis of a diverse set of easy-to-measure survey items which relate to a household's wealth. Components cumulatively measure asset ownership (such as televisions and bicycles), housing quality (e.g., construction materials), and access to utilities such as water and sanitation facilities [12]. We used this wealth index as it was reported in the DHS records for each cluster.

Education. The most reliable measure of household education captured by the DHS is the highest level of education attained by the survey respondent. In most cases, this respondent was a head of household (>95% of the time for Rwanda, Haiti, and Nepal). Our "education index" therefore roughly measures the typical level of education of the head of a household within a cluster. This is reported as one of four levels: "No education", "Primary", "Secondary", and "Higher". Any member below the lower age limit for the education questions was classified in the "No education" category. We assign an ordinal value between 0 and 3 to these categories, from "No education" to "Higher". Within each DHS cluster, the education index was computed as the average of this ordinal value of education level across all households.

Access to Water. The DHS captures the time it would take for a household member to reach a source of drinking water, in minutes. This value ranges from 1 minute to 500 minutes. For all the households that use piped-water, rainwater, have water-well in residence, or use bottled water, we assign them a value of 0.

Health Indices. The DHS captures rich information on health outcomes. We focus on three indices: the average hemoglobin level adjusted by altitude of all household adults (in g/dl), the average

Dataset	Rwanda	Nigeria	Haiti	Nepal
Panel A: Demographic and Health Survey Data (Source: DHS Program)				
Years Collected	2010	2013	2012	2011
Number of clusters	492	896	445	289
Number of households	12,540	38,522	13,181	10,826
Panel B: Satellite Nightlights Data (Source: NOAA DMSP-OLS)				
Number of $1x1km^2$ pixels	29,627	847,958	33,490	195,618
Low intensity pixels	28,579	793,643	29,731	186,360
Medium intensity pixels	955	51,390	3,592	9,133
High intensity pixels	93	2,925	167	125
Panel C: Daytime Satellite Imagery (Source Google Maps API)				
Area covered (km^2)	26,338	923,768	27,750	147,181
# images downloaded	29,627	1,036,956	33,490	195,618
# images in training set	26,665	112,425	30,141	176,057

Table 1: Primary datasets used in this study

body-mass-index (BMI) of the household's females, and the average number of mosquito bed nets per household.

Anthropometric Indices. These indices include the average height-for-age percentile, the average weight-for-age percentile, and the average weight-for-height percentile. Each is calculated by taking the mean of all individual childrens' measurements.

Electricity and Phones. Last, we study the percentage of households with access to electricity, and the percentage of households with one or more mobile phones. To calculate these values, we compute the fraction of households in a cluster who respond affirmatively to the question.

2.2 Daytime satellite images

We downloaded millions of satellite images from the Google Static Maps API (zoom level 16, pixel resolution 2.5m), which provide the input to our algorithms. These images cover the land area of Rwanda, Nigeria, Haiti, and Nepal, and are summarized in Panel B of Table 1. The size of each downloaded image is set to 400 × 400 pixels in order to match the land area covered by a single pixel of nighttime lights data, which are much lower resolution and typically cover 1 km^2 (more details below). We used shapefiles from the GADM database to determine the boundaries of each country [1].

2.3 Nightlight luminosity data

We obtained the NOAA nighttime light images from the DMSP-OLS website[1] for both F16 and F18 satellites in 2010, 2011, 2012, and 2013, to match the years of the DHS data for Rwanda, Haiti, Nepal, and Nigeria, respectively. The NOAA nighttime light intensity data includes a discretized luminosity level from 0 to 63, with 0 being the darkest pixel. We average the two satellites F16 and F18's data using ArcGIS software. Before using the nightlights data in the convolutional neural network, we remove gas flares[2] and then convert each pixel of light intensity into one of the three light intensity classes: low (0), medium (1) and high (2). Following Jean et al., we assigned the low intensity category to pixel values from 0 to 2, the medium intensity category to pixel values from 3 to 34, and the high intensity category pixel values from 35 to 63.

3 METHODS

To predict development indicators from daytime satellite images, we leverage a transfer learning process originally introduced by Xie et al. [26] and refined by Jean et al. [15]. This process is summarized in Figure 1. First, we calculate the average ground truth "welfare" of each geographic cluster in a country, where a cluster is defined by the DHS and is roughly equivalent to a rural village or urban neighborhood, and welfare is defined using a variety of different development indicators (such as wealth, education, and so forth) that are collected in the DHS household survey. Second, we compute satellite-based "features" for each cluster, by using a convolutional neural network (CNN) to extract features from satellite imagery covering the region in and around the cluster. This CNN is pre-trained on ImageNet, and then fine-tuned to predict categories of nighttime light intensity from daytime satellite images. Finally, we use a ridge regression model to learn the functional mapping from satellite features to development indicators at the cluster level. The tuned CNN and linear model can then be used to predict development indicators given arbitrary daytime satellite images.

3.1 "Ground truth" estimates of village welfare

Each household in the DHS data is assigned to a "cluster," or a community. The DHS data includes 492 clusters for Rwanda, 445 for Haiti, 289 clusters for Nepal, and 896 for Nigeria. The value of each development indicator (see Section 2.1) for a cluster is computed

[1] https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html

[2] Gas flares are demarcated in a separate shapefile available from NOAA [17].

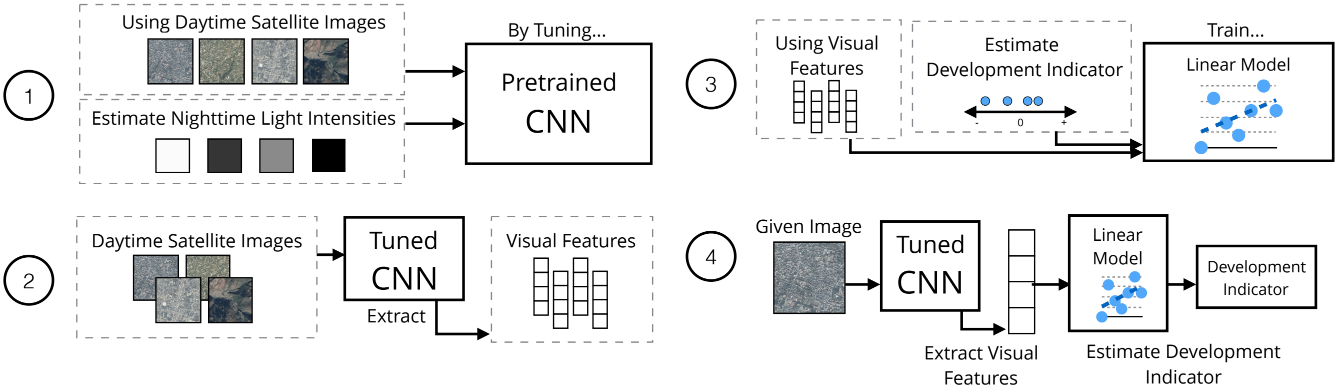


Figure 1: The transfer learning process used to predict development indicators from daytime satellite images. (1) A pre-trained CNN is tuned to predict nighttime light intensity; (2) High-level visual features are extracted from the top layers of the tuned CNN; (3) A linear model is trained to estimate a development indicator using ridge regression. (4) Given an arbitrary image, we can predict the development indicator by feeding the extracted visual features into the trained linear model. We tune a separate CNN for each country, and train a linear model to predict each development indicator for each country.

as the average across households within each geographic cluster, weighting each household by sample weights provided in the DHS survey to make the measurement representative. We discard 7 clusters in Nigeria and 8 clusters in Haiti that were recorded with a latitude and longitude of (0, 0), in addition to 59 other clusters in Nigeria that are affected by gas flares.

3.2 Extracting features from satellite imagery

For each country, we trained a neural network to predict nighttime light intensity using daytime satellite images. We started with a CNN with VGG16 architecture that had been pre-trained to recognize objects from the ImageNet dataset. Then, for each country, we fine-tuned the CNN to predict discretized nighttime light intensity from input daytime images. Following prior work [15, 26], we aimed to "teach" the network to recognize high-level visual features that were correlated with economic well-being.

For Rwanda, Haiti, and Nepal, we split the full set of images for each country into a training set (90% of the images) and a validation set (10%). Similarly to Xie et al. [26], we up-sampled images of high light intensity (classes 1 and 2) to enable balanced representation across classes. In our case, we up-sampled until all three classes of light intensity had the same number of training samples. We observed poor performance classifying daytime images with high nighttime light intensity without introducing this balance. For Nigeria, we sampled 60,000 images from each class (up-sampling classes 1 and 2) before splitting this sample into a training and validation set. We chose the trade-off of training for dozens of epochs over a subset of Nigeria's image data, over a few epochs over all one million images. We note that we were able to successfully replicated Jean et al.'s performance on Nigeria, despite training with only a sample of Nigeria's images.

Then, we removed the fully-connected top layers from the CNN, and added randomly-initialized, fully convolutional top layers as described in Xie et al. [26]. The top layers were retrained to predict nighttime light intensity. We determined that an initial learning rate of 2×10^{-3}, momentum of 0.9, and batch size of 100 were sufficient for this first round of tuning for the upper layers. In each epoch, we trained on the full training set. The learning rate was decreased by a factor of two whenever the validation loss stopped decreasing between epochs; tuning finished once the learning rate dropped below 10^{-5}. The hyperparameters for this stage were chosen heuristically and imprecisely, as we only wanted to initialize the top layers to more reasonable defaults than random values, and the weights in these layers were expected to change further in the following fine-tuning stages.

Next, we fine-tuned the full network, replicating the process described in prior work [15, 26]. We augmented the image data by mirroring each image horizontally. Like Xie et al., we began tuning with a initial learning rate of 10^{-6}. All convolutional layers had L2 regularization of $\lambda = 5 \times 10^{-4}$, following the hyperparameter settings from the original VGG paper [8]. We trained with a momentum of 0.9 and, due to hardware constraints, a batch size of 16. While Xie et al. tuned the model for 300,000 iterations, we followed the approach from the original VGG paper [8], decreasing the learning rate by a factor of ten whenever the validation loss stopped decreasing. Training stopped once the learning rate dropped below 10^{-10}. Whenever the learning rate decreased, training began again from the model with minimum validation loss from the previous learning rate, to avoid over-fitting.

In this way, we fine-tuned the CNN to predict nighttime light intensity for both Haiti and Rwanda. While this was relatively quick for Haiti and Rwanda, we used an alternative training process to scale to the larger corpora of images from Nepal and Nigeria: we skipped data augmentation, increased batch size to 32, and froze almost all layers in the CNN, tuning only the top layers and the last block of convolutional layers. Otherwise, the hyperparameters remained the same. For Nigeria and Nepal, we stopped the training

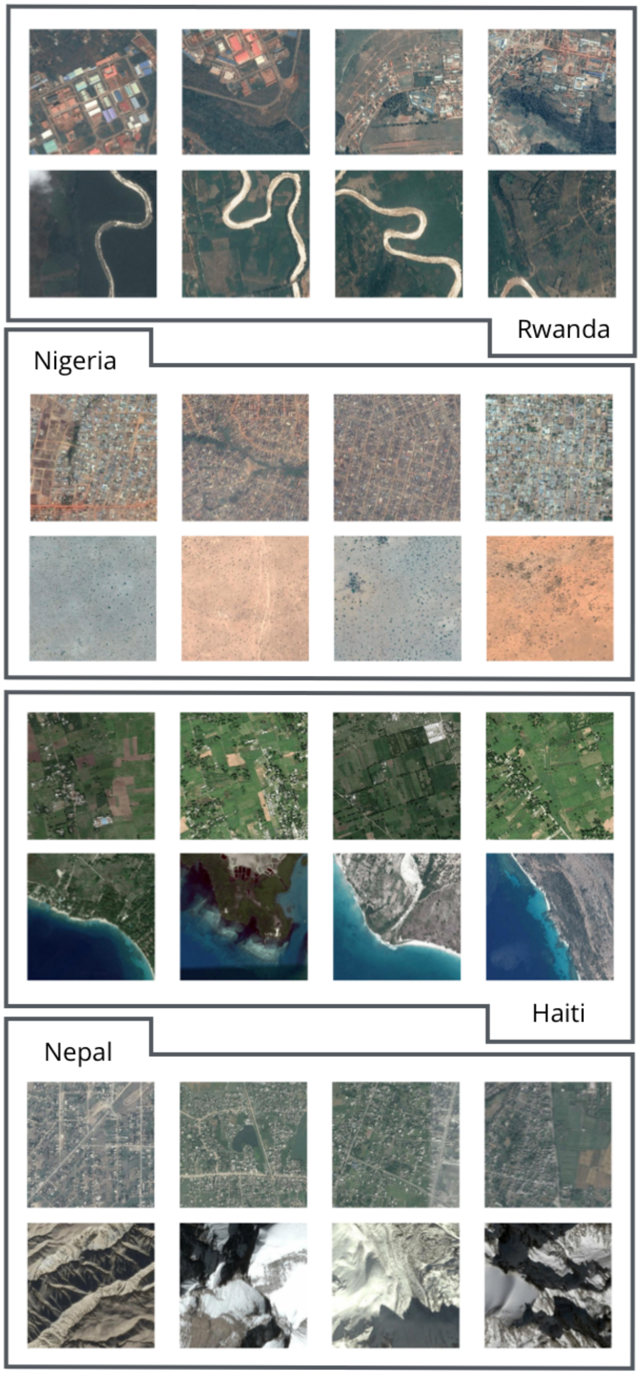


Figure 2: Satellite images and features. After fine-tuning, the CNNs learned to recognize visual features indicating human presence, and even notable geographical features of each region. Each row in the above subfigures represents a set of images that maximally "activate" filters from block 5 of each country's CNNs.

after 200,000 batches (about two and a half days of runtime), as the learning rate had not yet dropped below 10^{-10}.[3]

After tuning the CNN, it can be thought of as a function that maps raw satellite images to a set of "visual features," by passing images in as input and extracting the 16,384-dimensional vector of activations in the top layer. These features are optimized by the CNN to distinguish between regions of low, medium, and high nighttime light intensity. Some intuition for these features is shown in Figure 2. In the figure, we select a set of images that maximally activate one of the convolutional layers in block 5 of the network.

3.3 Mapping satellite features to development

Through the steps described above, the DHS data can be used to construct development indicators at the level of the cluster (our "target" variable), and the CNN can produce 16,384-dimensional feature vectors from each satellite image. We merge these two datasets by first identifying all satellite images in a 10×10 cell of images ($10km^2$) surrounding each cluster centroid. Each image is converted into a vector, and then, following prior work [15], the 100 image feature vectors from each cluster are averaged into a single vector for each cluster (our ”predictor” variables).

The final step is to learn the functional mapping from the cluster-level feature vector to the cluster-level development index. Per Table 1, the countries we analyze have between 289 and 896 such cluster-level observations. We perform this modeling using a regularized (ridge) regression model with five-fold cross validation. Regression hyperparameters (i.e., the ridge coefficient) are chosen in an internal (also five-fold) cross-validation loop. This model is fit separately in each country, and for each development indicator. We report the performance of each model as the average R^2 across all held-out folds in the outer cross-validation loop.

4 RESULTS

4.1 Wealth predictions in sub-Saharan Africa

Our first result involves replicating previously published results, to ensure that subsequent results properly extend the current state of the art. For this exercise, we applied the CNNs trained using the transfer learning technique described in [15] and [26] to estimate the asset-based wealth of villages (DHS clusters) in two countries that were analyzed in prior work. From the five countries studied in [15] (Malawi, Nigeria, Rwanda, Tanzania, Uganda), we focus on Nigeria and Rwanda, which have very different socio-demographic profiles.

Figure 3 illustrates our ability to replicate prior results in Rwanda (the two graphs on the left) and Nigeria (the two graphs on the right). In each country, we show the original result published in *Science* (reprinted with permission from AAAS) as well as the results we achieve when reproducing the results from scratch. Each scatter-plot contains one point for each of the clusters in the DHS

[3] We made this simplification to reduce the high computational cost involved in fully re-training the CNN. To validate that this simplified tuning process—tuning only the final convolutional layers of the CNN—did not significantly impact performance, we show in Appendix Figure 8 that fine-tuning the block 5 convolutional layers without data augmentation yields comparable results to the full training procedure described by Jean et al. [15]. On the same test sets, we saw comparable performance (R^2 within a few hundredths or less) in Haiti and Rwanda for estimating wealth, level of education, and access to water.

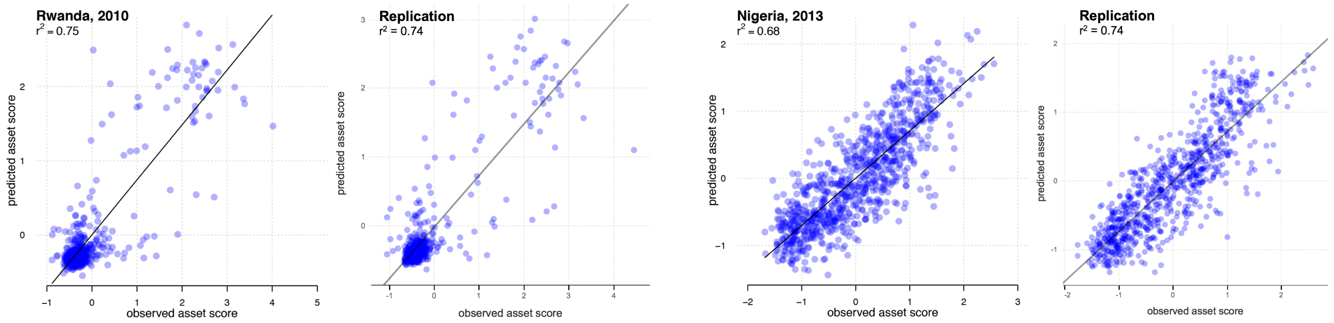


Figure 3: Predicting wealth from satellite imagery in sub-Saharan Africa. We replicated Jean et al.'s finding that a DHS asset index can be predicted with an R^2 value of about 70%. In each pair of plots, the plot on the left shows the per-cluster expected and predicted asset scores from Jean et al. The plots on the right of each pair shows our model's performance for the asset index on the same clusters from Rwanda and Nigeria. The figures on the left are reproduced from Jean et al., "Combining satellite imagery and machine learning to predict poverty", *Science* [15]. Reprinted with permission from AAAS.

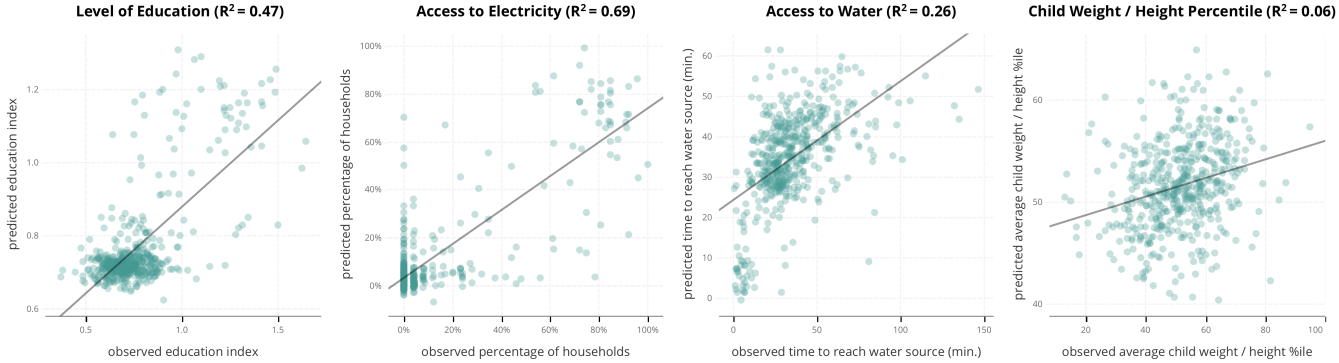


Figure 4: Predicting level of education, access to electricity and water, and health outcomes in Rwanda. Prediction performance varies widely for different development indicators. In Rwanda, the average R^2 for estimating development indicators ranged from a high of 0.74 for an asset-based wealth index, to a low of 0.06 for predicting the average weight / height percentile for children.

survey, with the x-axis indicating the actual wealth of the cluster (as measured in the DHS), and the y-axis indicating the wealth predicted through the transfer learning algorithm. Visually, our results are very similar to those in prior work. Quantitatively, in Rwanda, our model can reconstruct village wealth with an R^2 of 0.74 (vs. the R^2 of 0.75 reported in [15]); in Nigeria, we achieve similar performance of R^2=0.74 (vs. R^2=0.68 reported in [15]).

4.2 Wealth predictions outside Africa

Previously published results on using deep learning to predict wealth have only focused on five countries in sub-Saharan Africa. There, the predictive models achieved goodness of fit ranging from 0.55 to 0.75 (Malawi: 0.55; Tanzania: 0.57; Nigeria: 0.68; Uganda: 0.69; Rwanda: 0.75). [15] also present evidence that their model can "travel well across borders," i.e., that a model trained in one of those five countries can be applied with reasonable success to estimating wealth in a different one of those five countries (figure 5 in [15]). However, prior work does not address whether this approach can be applied outside of the sub-Saharan African context, where the relative homogeneity of the geographic environment might be uniquely well-suited to this method.

Our first novel result is to test the ability of this same modeling approach to generalize outside of sub-Saharan Africa. Here, we focused on Haiti and Nepal, two geographically diverse contexts small enough to facilitate rapid iteration of our model training and tuning procedures.[4] On these countries, we find that the model's ability to predict wealth was in one case slightly lower than previous results in sub-Saharan Africa (Haiti: R^2=.51%) and in another case comparable (Nepal: R^2=0.64). This suggests that these techniques are not uniquely appropriate to sub-Saharan Africa. At the same time, there may be certain specific countries where these methods perform quite poorly. Thus, in ongoing work we are applying this approach to a much larger set of countries across the globe (initial results have not revealed any countries with markedly lower performance).

[4] In a country as large as Nigeria, it takes several weeks to download and process the raw satellite imagery, and several days of GPU computing to train the CNN.

4.3 Generalizing to other measurements

Our second set of results test the ability of the original transfer learning approach to generalize to the prediction of development indicators other than the asset-based wealth index. Here, we find that the model cannot predict any of the other development indicators as accurately as it predicts wealth, and that for certain indicators the performance is no better than a random guess. These results are produced when applying the exact same formula (described in Section 3) to a different development indicator; we discuss later whether modifications to this formula might improve these results.

Figure 4 illustrates this point in Rwanda. Going from left to right, we show the model's ability to measure levels of education, electrification, access to water, and child weight-for-height index. In all cases the performance is lower than the performance for the wealth index (R^2=0.74, the second scatter-plot in Figure 3). For both education and electrification the performance is still reasonable (R^2=0.47 and R^2=0.69, respectively), but degrades significantly when predicting the access to water (R^2=0.26) and the average children weight-for-height index (R^2=0.06).

Outside of Rwanda, the results for non-wealth based measures of development degrade more rapidly. This evidence is summarized in Figure 5, which shows the performance of the model in predicting eleven different measures of human development in each of the four countries. In the figure, each box-and-whisker plot shows the distribution of test R^2 values from the outer loop of cross-validation. The full set of results from Rwanda are shown in the top set of plots. As described earlier, performance is best for wealth and electricity, and worst for health-related indices. Figure 6 collapses these results by indicator, making more explicit the range of average R^2 values for each indicator in each country.

Across all countries and development indicators, we observe a range of R^2 performance from -0.02 (equivalent to a random guess) to 0.73 (highly accurate). Within this range, several patterns emerge. First, in all countries, the model achieves the best performance in predicting the asset-based wealth index. This was the headline result from prior work, and it appears to be quite robust. As we discuss below, there are several factors that might cause this approach to work best for wealth prediction.

Second, there appears to be a separate group of development indicators—health and anthropometric indices—that are consistently difficult to predict in each country. This group includes the average child weight-to-height percentile, for which the maximum R^2 of any country is 0.11. This group also includes hemoglobin level, and the average child height and weight percentiles. For each of these indicators, the R^2 is consistently lower than those reported for the asset index in prior work. However, we note one surprising outlier in this group: female body mass index can be predicted with R^2 between 0.31 and 0.47.

Third, we note a group of indicators where performance is modest, and where additional refinement may hold promise. This includes predicting levels of education (R^2 between 0.47 and 0.64), access to electricity (R^2 between 0.24 and 0.69), and mobile phone ownership. This may be in part due to the fact that all three are closely correlated with wealth (and in case of the latter two, may even be mechanically factored in to the wealth index).

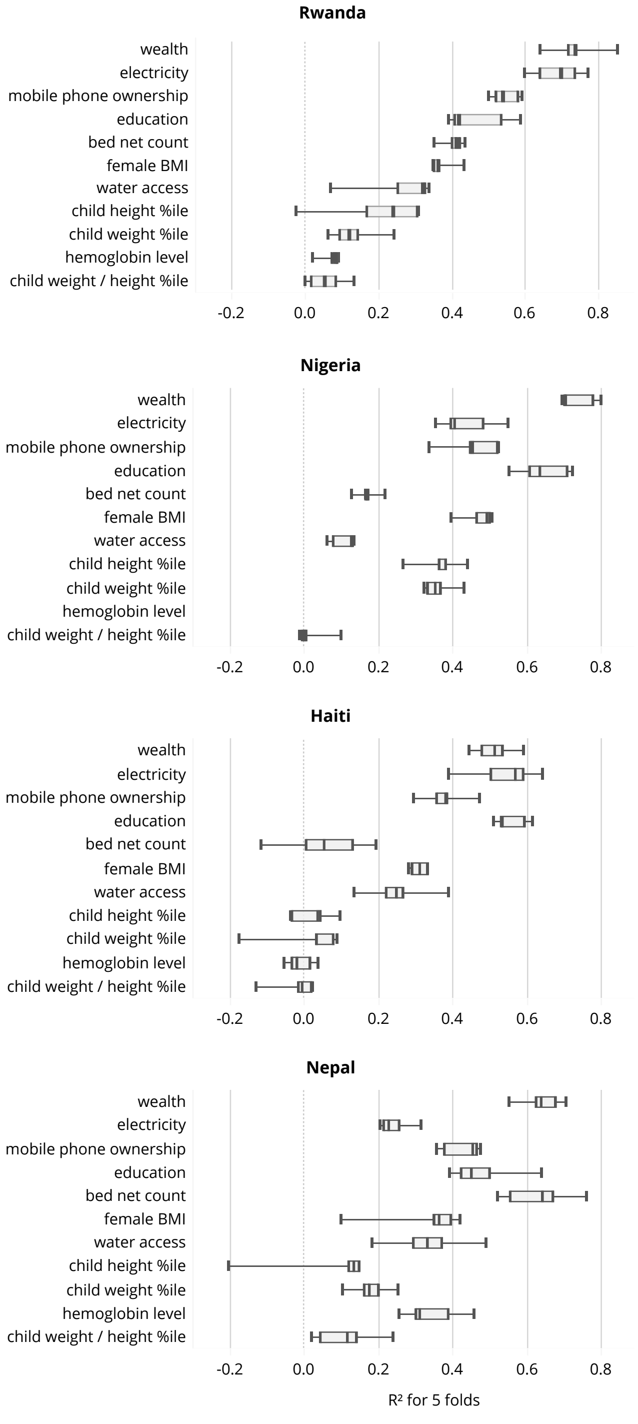


Figure 5: Generalizing the approach to other indicators of development in other countries. Boxes and whiskers mark the range of R^2 values from five-fold cross-validation. Indicators are ordered from highest to lowest median R^2 in Rwanda.

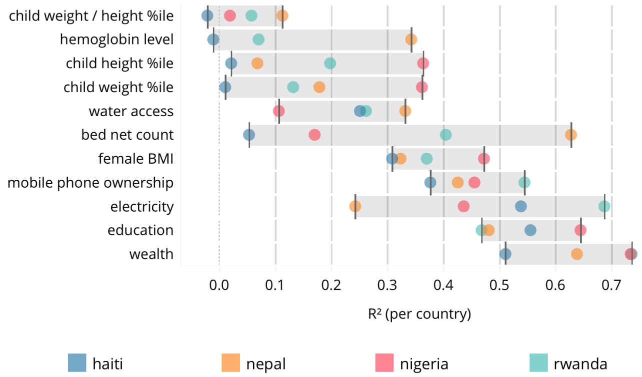


Figure 6: Which development indicators are hardest to predict? Some indicators are difficult to predict from visual features extracted from satellite images (average child weight / height percentile). Others can be predicted consistently well in every country (wealth, education). For some indicators, prediction accuracy varies greatly between countries (household bed net count). Colored dots show the average R^2 for predicting the development indicator from satellite images in each country. Gray bars highlight the range from minimum to maximum R^2.

Fourth, we note that for several indicators (access to electricity, bed net count, water access), the average R^2 varies widely across countries (see Figure 6). In the most extreme case, bed net count is predicted with an R^2 of only 0.05 in Haiti, and an R^2 of 0.63 in Nepal. These high variance situations suggest there might be regional variations in the landscapes and human structures that correspond to the indicators. In some cases (as with water access) there is also high variance within a country, between different test folds in cross-validation. This may be indicative of a high level of geographic heterogeneity within the country, leading to varying performance when certain types of regions are randomly included or excluded from the training set.

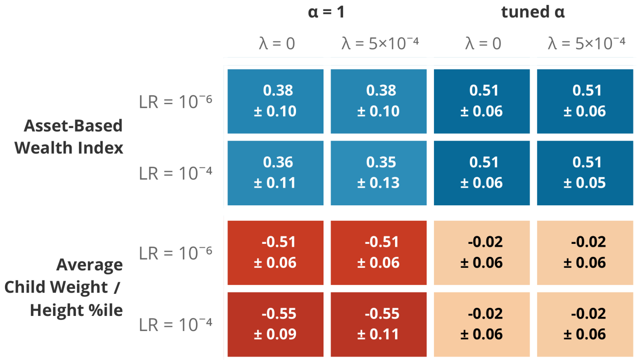
Finally, it appears there is no one country in which prediction is uniformly better for all indicators. In general, however, and given the limited sample of 11 indicators from 4 countries, performance is better in the sub-Saharan countries than in Haiti and Nepal. This is the case for the wealth index, mobile phone ownership, and female BMIs. There are also two outlying indicators, hemoglobin level and bed net count, for which the model predicts better in Nepal by more than 0.20 relative to the next country.

4.4 Model tuning and optimization

For researchers and practitioners interested in using satellite-based measurements of development as inputs into current and future projects, an important lesson we have learned in fitting the above models is that prediction accuracy can be sensitive to hyperparameter choice in unexpected ways. In particular, we found that models can be quite brittle, and that the fact that a model performs well on a given metric in a given country does not mean that it will necessarily work out of context.

In particular, in our initial set of experiments, we used a simplified training procedure that relied heavily on the default settings of the CNN and ridge regression algorithm. This model predicted the wealth index in Rwanda nearly as well ($R^2 = 0.71$) as the original results published in *Science* ($R^2 = 0.75$) [15], and as the optimized results described above ($R^2 = 0.74$). However, even for wealth, this model performed quite poorly in other countries ($R^2 = 0.35$ for Nepal and $R^2 = 0.53$ for Nigeria).

What could explain these differences? Key differences between our initial and final model include training using a more aggressive learning rate (10^{-4} vs. 10^{-6}), a larger training set of images (10,000 images per class nighttime light intensity vs. 60,000 or all), and carefully tuning the ridge regression regularization parameter (vs. using the default value of 1). To better understand these discrepancies, we repeated tuning for Haiti using all combinations of these parameter values, thus populating a 2x2x2 grid of possible hyperparameters. This grid is shown for the wealth index and for the child weight / height index in Figure 7.[5]



		α = 1		tuned α	
		$\lambda = 0$	$\lambda = 5\times10^{-4}$	$\lambda = 0$	$\lambda = 5\times10^{-4}$
Asset-Based Wealth Index	LR = 10^{-6}	0.38 ± 0.10	0.38 ± 0.10	0.51 ± 0.06	0.51 ± 0.06
	LR = 10^{-4}	0.36 ± 0.11	0.35 ± 0.13	0.51 ± 0.06	0.51 ± 0.05
Average Child Weight / Height %ile	LR = 10^{-6}	-0.51 ± 0.06	-0.51 ± 0.06	-0.02 ± 0.06	-0.02 ± 0.06
	LR = 10^{-4}	-0.55 ± 0.09	-0.55 ± 0.11	-0.02 ± 0.06	-0.02 ± 0.06

Figure 7: Sensitivity to hyperparameters. We varied the hyperparameters of initial learning rate (LR), kernel weight regularization (λ), and ridge regression regularization (α) to see how they impacted prediction accuracy. For the tested configurations, the CNN was robust to more aggressive learning rates (LR = 10^{-4}) and unpenalized kernel weights ($\lambda = 0$). However, using default regression regularization settings caused major performance degradation.

Prediction performance was resilient to the tested changes to the CNN configuration: changing the starting learning rate from 10^{-6} to 10^{-4} affected average R^2 by no more than 0.04, and this difference only occurred when the ridge regression α was initialized as its default value. Disabling kernel weight regularization had no detectable impact on prediction accuracy. However, prediction accuracy became much more volatile when the ridge regression configuration changed. When the ridge regression regularization parameter was set to its default value rather than a tuned value found through cross-validation, the observed R^2 dropped by at least 0.13 for the asset-based wealth index, and by about 0.50 for the measurements of the average child weight to height percentile.

It's sensible to think one would already know to tune a regularization parameter when providing input to a linear regression with

[5] For these trials, we used the same tuning process as we described in Section 3.2, with the modification that training was stopped after the learning rate dropped below 10^{-8}. This change was made to expedite the analysis. We find empirically that despite this earlier stopping point, the models trained during these tests performed similarly to the model described in Section 4.3.

tens of thousands of factors. At the same time, it might not be obvious that a CNN would face substantial performance degradation in a transfer learning task, especially when it achieved high validation accuracy over dozens of epochs. While CNN tuning appears to be resilient to a number of reasonable hyperparameter choices, we caution those adapting this technique to be aware of the major impact seemingly small adaptation choices can have on the overall success of the transfer learning process.

5 DISCUSSION

Our results in this paper have shown that while satellite imagery and machine learning may provide a powerful paradigm for estimating the wealth of small regions in sub-Saharan Africa, the same approach does not trivially generalize to other geographical contexts or to other measures of human development. In this assessment, it is important to emphasize what we mean by "trivially," because in truth the point we are making is somewhat circumspect. Specifically, what we have shown is that the exact framework—of retraining a deep neural network on night-lights data, and then using those features to predict the wealth of small regions in sub-Saharan Africa—cannot be directly applied to predicting arbitrary indicators in any country with uniformly good results.

We believe that this is an important point to make because absent empirical evidence to the contrary, it is likely that policymakers eager to gain quick access to micro-regional measurements of development might be tempted to do exactly what we have done in this paper, without paying careful attention to the thorny issues of generalizability that we have uncovered in this analysis. It is not our intent to impugn the potential for related approaches to provide important new methods for measuring development, but rather to say that such efforts should proceed with caution, and with careful validation.

Our results showed that while some indicators like wealth and education can be predicted reasonably well in many countries, others development indicators are much more brittle, exhibiting high variance between and within countries, and others perform poorly everywhere. We thus find it useful to distinguish between two possible reasons why the current approach may have failed to generalize to these measures of development. First, it may be that this exercise is fundamentally not possible, and that no amount of additional work would yield qualitatively different results. Second, it is quite possible that our investigation to date has been not been sufficiently thorough, and that more concerted efforts could significantly improve the performance of these models. We discuss each possibility in turn.

5.1 Fundamental limitations

There are several possible reasons why it might be fundamentally impossible to use satellite imagery to accurately measure certain aspects of human development. These include:

Insufficient "signal" in the satellite imagery. Our overarching goal is to use information in satellite images to measure different aspects of human development. The premise of such an approach is that the original satellite imagery must contain useful information about the development indicator of interest. Absent of such a signal, no matter how sophisticated our computational model, the model is destined to fail.

The fact that *wealth* specifically can be measured from satellite imagery is quite intuitive. For instance, looking at the photos in Figure 2, there are visual features one might expect correlate with wealth—large buildings, metals roofs, nicely paved roads, and so forth. It may be the case that other measures of human development cannot be seen from above. For instance, it may be a fundamentally difficult task to infer the prevalence of malnutrition from satellite imagery, if the regions with high and low rates of malnutrition appear similar, even though we hypothesize that these indices should correlate with wealth index [24]. We were, however, surprised by the relative under-performance of models designed to predict access to drinking water, as we expected the satellite-based features to capture proximity to bodies of water, which in turn might affect access to drinking water.

(Over-) reliance on night-lights may not generalize. Our reliance on night lights might help explain why some indicators were predicted less successfully in some countries than others. An example in our study includes Nepal, where the accuracy in predicting access to electricity was much lower ($R^2 = 0.24$) than in the other countries (R^2 = 0.69, 0.44, and 0.54 in Rwanda, Nigeria, and Haiti, respectively). This may be partly due to the fact that Nepal has a very low population density (half as dense as Haiti and Rwanda) and very high levels of electrification (twice as high as Haiti, Rwanda, and Nigeria) [2]. If the links between electrification, night-lights, and daytime imagery are broken in Nepal, we would expect our modeling approach to fail. More generally, we expect that when a development indicator does not clearly relate to the presence of nighttime lights, it may be unreasonable to expect good performance from the transfer learning process as a whole.

Deep learning vs. supervised feature engineering. In this paper, we have focused explicitly on using the deep/transfer learning approach to extracting information from satellite images. While powerful, it is also possible that other approaches to feature engineering might be more successful than the brute force approach of the convolutional neural network. For instance, Gros and Tiecke [13] have recently shown how hand-labeled features from satellites, and specifically information about the types of buildings that are present in each image, can be quite effective in predicting population density. Labeling images in this manner is resource intensive, and we did not have the opportunity to test such approaches. However, we believe that careful encoding of the relevant information from satellite imagery would likely bolster the performance of specific prediction tasks.

5.2 Possible shortcomings

The issues discussed above represent a set of potential limitations that are inherent to predicting human development with satellite data and deep learning. In addition, it is possible—and perhaps quite likely—that the overarching approach *could* work, given more concerted and focused analysis. Here, we discuss a few salient improvements to the experimental design, data collection, and deep learning architecture that would represent useful next steps for future work.

Measurement error. One difficulty we encountered was the high levels of measurement error in our "ground truth" data. In the case of the DHS, for example, noise is intentionally added to the GPS coordinates of each household, to preserve the privacy of individual respondents. Since we rely on these coordinates to match surveys to satellite images, this introduces classical measurement error in our response variable.

This error is also likely different for each measure of development that we tested. For instance, the wealth index is a composite index, calculated as the first principal component of a large number of questions related to household characteristics and asset ownership. Such a composite index likely smooths out considerable noise in the individual asset questions. By contrast, most of our other development indicators are derived from a single survey question, and may thus have a lower ratio of signal to noise.

Mis-specification. As we have shown, the effectiveness of this approach depends on careful tuning of the algorithms used to derive development estimates from satellite imagery. It is even possible—though we think unlikely—that a different set of tuning parameters could be used to estimate anthropomorphic outcomes (the lowest-performing indices) from the same input data. To more conclusively investigate such possibilities, we think an important, albeit computationally intensive, next step involves systematically exploring the parameter and hyperparameter space of the machine learning algorithms. In our case, we relied heavily on the TensorFlow implementation of VGG16, as it provided considerable functionality. But this functionality comes at the cost of flexibility, and careful tuning might improve predictive performance. Examples of design decisions to explore include: the type of CNN (we use VGG16), supervised learning algorithm (ridge regression), cross-validation procedure (inner/outer 5-fold), image resolution (zoom 16) and augmentation (mirroring), image labels (nighttime lights) and discretization (3 bins), batch size (32), CNN parameters (learning rate, regularization, momentum), and many others.

Countries of focus. A final issue worth highlighting is that our assessment of the ability of the original model to generalize to other contexts is based on an evaluation of Haiti and Nepal. These countries were chosen because they are dissimilar to the original sub-Saharan nations, but also because they are relatively small nations where the required data and computational resources were more manageable. However, our conclusions about geographic generalizability should be qualified by the understanding that Haiti and Nepal are by no means representative of all other developing countries, and that additional testing is required before affirmative claims about global generalizability can be made.[6]

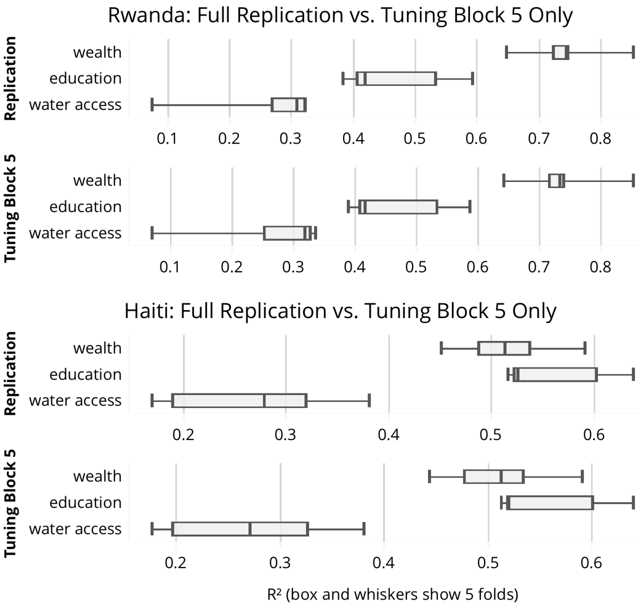


Figure 8: Speeding up training by tuning block 5 only. We achieved similar prediction performance tuning only the CNN's top layers and fifth block of convolutional layers, as when we replicated the training process described in [15]. The distribution of R^2 values across folds were strikingly similar for the two models for each of several development indicators in Rwanda and Haiti.

6 CONCLUSION

This project presents a preliminary investigation of the generalizability of satellite-based methods for estimating human development. After replicating prior work that established the potential for such methods to predict asset-based wealth in Rwanda, we show that the same approach cannot be trivially translated into predicting other "softer" development outcomes (such as health outcomes and access to clean drinking water) with the same accuracy in other countries (specifically, Haiti and Nepal). We provide some discussion of how this failure to generalize could be caused by fundamental constraints of the satellite-based approach, as well as specific issues with implementation. We also outline several compelling next steps for research in this space.

Broadly, we remain optimistic that future work using novel sources of data and new computational algorithms can engender significant advances in the measurement of human development. However, it is imperative that such work proceeds carefully, with appropriate benchmarking and external calibration. Promising new tools for measurement have the potential to be implemented widely, possibly by individuals who do not have extensive expertise in the underlying algorithms. Applied blindly, these algorithms have the potential to skew subsequent policy in unpredictable and undesirable ways. We view the results of this study as a cautionary example of how a promising algorithm should not be expected to work "off the shelf" in a context that is significantly different from the one in which it was originally developed.

REFERENCES

[1] Global Administrative Areas. 2013. GADM databases of Global Administrative Areas. http://www.gadm.org/. (2013).

[2] World Bank. 2017. Access to electricity (percent population). https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS. (2017).

[3] World Bank. 2017. Population living in slums (% of urban population). https://data.worldbank.org/indicator/EN.POP.SLUM.UR.ZS?locations=NP-HT-NG-RW. (2017).

[6] Haiti, for instance, had the weakest performance for 8 out of 11 development indicators. With its disproportionately urban population (60% of total, with 75% living in slums [3]), it may be representative of countries that pose unique challenges for learning meaningful visual signals of development.

[4] Tara Bedi, Aline Coudouel, and Kenneth Simler. 2007. *More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions.* World Bank Publications.

[5] Daniel J. Benjamin, Ori Heffetz, Miles S. Kimball, and Nichole Szembrot. 2014. Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference. *The American Economic Review* 104, 9 (2014), 2698–2735. http://www.jstor.org/stable/43495330

[6] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350, 6264 (Nov. 2015), 1073–1076. https://doi.org/10.1126/science.aac4420

[7] Joshua Evan Blumenstock. 2016. Fighting poverty with data. *Science* 353, 6301 (Aug. 2016), 753–754. https://doi.org/10.1126/science.aah5217

[8] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* (2014).

[9] Xi Chen and William D. Nordhaus. 2011. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences* 108, 21 (May 2011), 8589–8594. https://doi.org/10.1073/pnas.1017031108

[10] Angus Deaton. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* World Bank Publications.

[11] Deon Filmer and Lant H. Pritchett. 2001. Estimating Wealth Effects Without Expenditure Data Or Tears: An Application To Educational Enrollments In States Of India*. *Demography* 38, 1 (Feb. 2001), 115–132. https://doi.org/10.1353/dem.2001.0003

[12] United States Agency for International Development. 2013. Demographics and Health Survey. Standard Recode Manual. (march 2013), 1–171. http://www.dhsprogram.com/pubs/pdf/DHSG4/Recode6_DHS_22March2013_DHSG4.pdf

[13] Andreas Gros and Tobias Tiecke. 2016. https://code.facebook.com/posts/1676452492623525/connecting-the-world-with-better-maps/. (2016).

[14] J. Vernon Henderson, Adam Storeygard, and David N. Weil. 2012. Measuring Economic Growth from Outer Space. *American Economic Review* 102, 2 (2012), 994–1028. https://doi.org/10.1257/aer.102.2.994

[15] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (Aug. 2016), 790–794. https://doi.org/DOI:10.1126/science.aaf7894

[16] Morten Jerven. 2013. *Poor numbers: how we are misled by African development statistics and what to do about it.* Cornell University Press.

[17] Matt Lowe. 2014. Night Lights and ArcGIS: A Brief Guide. *Emerg Infect Dis* (Jan. 2014). http://economics.mit.edu/files/8945

[18] Charlotta Mellander, Jose Lobo, Kevin Stolarick, and Zara Matheson. 2015. Night-Time Light Data: A Good Proxy Measure for Economic Activity. *PLoS One* 10, 10 (Oct. 2015), e0139779. https://doi.org/10.1371/journal.pone.0139779

[19] Sendhil Mullainathan. 2016. Satellite Images Can Pinpoint Poverty Where Surveys Can't. *New York Times* (April 2016). http://www.nytimes.com/2016/04/03/upshot/satellite-images-can-pinpoint-poverty-where-surveys-cant.html?_r=0

[20] United Nations. 2015. Transforming our world: the 2030 agenda for sustainable development. (2015), 1–35. https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf

[21] Abdisalan Noor, Victor Alegana, Peter Gething, Andrew J. Tatem, and Robert W. Snow. 2008. Using remotely sensed night-time light as a proxy for poverty in Africa. *Population Health Metrics* 6, 5 (Oct. 2008). https://doi.org/10.1186/1478-7954-6-5

[22] A. K. Sen. 1999. *Development as Freedom.* Oxford University Press.

[23] Jessica E. Steele, Pal Roe Sundsoy, Carla Pezzulo, Victor A. Alegana, Tomas J. Bird, Joshua Evan Blumenstock, Johannes Bjelland, Kenth Engo-Monsen, Yves-Alexandre de Montjoye, Asif M. Iqbal, Khandakar N. Hadiuzzaman, Xin Lu, Erik Wetter, Andrew J. Tatem, and Linus Bengtsson. 2017. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface* 14, 127 (Feb. 2017), 20160690. https://doi.org/10.1098/rsif.2016.0690

[24] Helga B. Urke, Torill Bull, and Maurice B. Mittelmark. 2011. Socioeconomic status and chronic child malnutrition: wealth and maternal education matter more in the Peruvian Andes than nationally. *Nutrition Research* 31, 10 (2011), 741–747. https://doi.org/10.1016/j.nutres.2011.09.007

[25] USAID. 2017. The DHS Program Overview. http://www.dhsprogram.com/Who-We-Are/About-Us.cfm. (2017).

[26] Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. 2016. Transfer learning from deep features for remote sensing and poverty mapping. *AAAI Conference on Artificial Intelligence,* arXiv:1510.00098v2 (Feb. 2016). https://arxiv.org/abs/1510.00098v2