Supplementary Materials

1 Methods and data

Our method relies on a transfer learning approach to extract image features from daytime satellite imagery using a convolutional neural network (CNN) (*15*). In an earlier paper that partially developed these methods (*15*), we focused on binary poverty classification for one country (Uganda). Our goal in the work described here is to build on these methods by extending the analysis to five countries, studying continuous measures of both consumption and assets, quantifying how models will do when trained on one country and applied to another, and comparing performance against alternative approaches to estimating these outcomes. Here we provide more detail about the methods and data used.

1.1 Image feature extraction

Deep learning methods, and convolutional neural networks in particular, have driven recent landmark advancements in computer vision, helped along by enormous datasets such as ImageNet, which contains millions of labeled training examples (*18*). The CNN model that we use is highly non-linear, has over 55 million parameters, and is sufficiently flexible to extract complex features from images, e.g., the presence/absence of a road. Since we have only several hundred data points for consumption or assets in each country to be used as labeled training examples, we cannot directly train a large CNN model to estimate these outcomes from satellite images—we simply do not have enough data. Additionally, the task of estimating economic well-being from satellite imagery is nontrivial for human non-experts, precluding the generation of additional labeled training data through crowdsourcing services such as Amazon Mechanical Turk.

To combat the data scarcity problem, we use a transfer learning method and train a fully-convolutional CNN model on the data-rich nighttime light estimation task. By solving this related proxy task, the model learns how to extract features that are also useful for the poverty estimation task. In previous work (*15*), we found that a multi-step approach outperforms simpler transfer learning methods that use imagery and nightlight information. For example, a simpler alternative would be to use an off-the-shelf CNN trained on ImageNet to extract image features that could then be used in conjunction with nightlights to predict poverty indicators (*15, 23*).

In the first step of the transfer learning approach, we fine-tune an 8-layer CNN model (VGG F) previously trained on the ImageNet dataset to estimate nighttime light intensity at various locations given the corresponding daytime satellite images (*24*). We treat this step of the transfer learning approach as a classification problem, with three nighttime light intensity classes obtained by fitting a mixture of three Gaussian distributions to the relative frequencies of the nighttime light intensity values. Intuitively, the three classes of nighttime light intensities correspond to low, medium, and high intensity. The three class distinctions were determined by observing the histogram of nighttime light intensities in our training set, which includes over 300,000 locations in Africa sampled near DHS (Demographic and Health Surveys) locations (note SM 1.4) (*25*). The histogram suggests that there are three dominant modes of nightlight intensities, and the Gaussian mixture model provides a principled method of of binning the data. Measured nightlight intensity values are integer values ranging from 0-63 (SM 1.5). This interval

is partitioned into a low class corresponding to near 0 nightlight intensity, a medium class corresponding to nightlight intensity roughly in the 3-34 range, and a high class corresponding to 35-63.

Since nighttime light data is available globally at a 1 km resolution (*26*), our inputs are 400×400 pixel daytime satellite images from Google Static Maps at zoom level 16, which roughly correspond to 1 square km areas. The Google Static Maps API does not directly provide temporal data, but each image has a small watermark denoting the year when the image was taken. The daytime satellite images used for this study were primarily collected from 2013 to 2015. As such, we use nightlight labels from 2013 to minimize the temporal differences between daytime image inputs and corresponding nightlight labels.

The model is trained with minibatch gradient descent with the momentum update scheme using the same hyperparameter settings as proposed in the original VGG paper (*24*). For complete details on the method and training process, readers can refer to previous work (*15*). After training the CNN model to predict nighttime light intensity, we use this learned model as a feature extractor for daytime satellite images by discarding the last layer of the CNN model, which is the nighttime light classification layer. For each household cluster, we use (up to) 100 input images that cover a 10 km by 10 km area centered around the cluster location. For Nigeria and Tanzania, the two largest countries in our sample by land area, we sample 25 evenly spaced images for each cluster. For the smaller countries, we fully tile the 10 km by 10 km area, sampling 100 images per cluster. We found that using this sparse sampling strategy for the larger countries did not affect the performance of the model. Since there is up to 5 km of jitter added in each direction for cluster locations, covering a 10 km by 10 km area ensures that the true location is seen somewhere in the input images, regardless of whether the area is rural or urban. Therefore, for each cluster we make (up to) 100 evaluations of the CNN model, resulting in 100 feature vectors. We then average these feature vectors to obtain one feature vector for the cluster, which is then used as input in regression models for estimating survey-based measures of either consumption expenditure or assets. Note that regularized linear regression models are used instead of another deep neural network because of the lack of sufficient data to train a complex model.

Spatial context is a key consideration when designing the architecture of the CNN model. Satellite images are usually large (in pixel dimensions), and important parts of the image are not necessarily in the center of the image, which is unlike typical images considered in computer vision problems. Before fine-tuning the pre-trained VGG F model, we convert the fully-connected layers of the CNN into convolutional layers followed by an average pooling layer. The modified architecture allows for input images of arbitrary size, whereas traditional CNN models have fully-connected layers at the end of the network with fixed input sizes that are dependent on the size of the input image. Intuitively, replacing these fully-connected layers with convolutional layers is an equivalent transformation for images of the same size, but allows the network to "slide" across larger images, producing multiple feature vectors by making multiple evaluations of the network across different parts of the image in an efficient manner using convolutions. The average pooling layer then averages these multiple feature vectors into a single feature vector that summarizes the input image, which can then be used for classification in the nighttime light task. Note that this averaging process is separate from the averaging process used in creating one feature vector per cluster; here, the network makes multiple evaluations of a single image via convolutions and averages the results to produce one feature vector per image.

1.2 Dimensionality reduction

The CNN extracts 4096-dimensional feature vectors from input satellite images. We find that using principal component analysis (PCA) to reduce the dimension retains much of the relevant feature information (e.g., reducing to the first 10 principal components retains ∼96-98% of data variation), so in several of our experiments where many trials need to be run (e.g., Fig. 4), we choose to first reduce the dimension of the feature vectors to reduce computational cost. Dimensionality reduction is noted wherever it is applied. The benefits are twofold—not only do we save on computation, but in the case of small training datasets, we also guard against overfitting by using a less complex model to represent the relationship between inputs and outputs. However, since we also use regularization in all of our linear regression models, we emphasize that dimensionality reduction is not strictly necessary for any of our experiments.

1.3 Filter activation maps

The CNN model parameters are represented as individual filters, which slide across the input daytime image and are specialized for looking for certain features, such as edges and lines in the earlier layers and more complex features such as buildings and roads in the later layers. We use the term "activating" to describe a filter that has detected something that it has been trained to look for in a certain part of the image. One way to visualize the CNN filters is to examine the set of images that activate each filter most strongly. For the visualizations in Fig. 2, we display a subset of maximally activating images and their corresponding activation maps from the fifth layer of the CNN model (*15*).

1.4 Survey data

We study two indicators of economic well-being. The first is consumption expenditure, as measured in the World Bank's Living Standards Measurement Surveys (LSMS). Annual consumption expenditure, or the total amount of money a household spends on consumption goods over 12 months, is the standard measure used in developing countries to classify households as poor or not poor. The LSMS surveys are conducted in many countries around the world, and we use the most recent LSMS surveys available in Africa: Nigeria 2012-13, Tanzania 2012-13, Uganda 2011-12, and Malawi 2013. LSMS surveys use a two-stage sampling design, in which enumeration areas (which we refer to as "clusters", roughly equivalent to villages in rural areas or wards in urban areas) are sampled throughout a country, with probability of sampling proportional to population, and then households are randomly sampled within each cluster.

The LSMS surveys we use are each part of a longitudinal study, with households tracked over the course of several years. As individual members split off and start their own households or otherwise move, new single household cluster coordinates are added to the dataset. To reduce noise, we use only cluster coordinates with multiple households when training and evaluating our models. Our resulting clusters contain between 2 and 20 households, with a mean of 10.9 and a median of 10.

For each survey, we averaged household consumption expenditures at the cluster level, and then used purchasing power parity exchange rates to convert measurements in each country to a common currency (2011 USD), allowing for direct comparison to the current World Bank global poverty line of $1.90 per capita per day. The distribution of consumption expenditures at the cluster level was roughly log-normal, so we decided to estimate log expenditures.

Cluster locations are reported as the average latitude and longitude of cluster households, plus some added noise (± up to 2 km in urban areas, up to 5 km in rural areas, and up to 10 km for a random 1% of clusters) to preserve the anonymity of survey respondents. Since we do not know the true location of each cluster, we sample daytime satellite imagery from a 10 km by 10 km square centered on the reported cluster location, and then average image features across the whole area. We cannot extract features without satellite imagery, so we restrict the clusters to those where at least 10% of the surrounding area has available Google Static Maps imagery (i.e., at least 10 images are available from within the 10 km by 10 km square centered on the cluster location). Our final LSMS dataset consists of 487 clusters in Nigeria, 405 clusters in Tanzania, 315 clusters in Uganda, and 204 clusters in Malawi, for a total of 1411 clusters across the four countries.

Our second measure of economic well-being is a household asset score taken from the Demographic and Health Surveys (DHS). For most developing countries in the world, DHS surveys collect nationally representative data on fertility, maternal and child health, HIV/AIDS, malaria, and nutrition. Many of these surveys also report a "wealth index", which is computed as the first principal component of survey responses for a set of questions about common asset ownership (e.g., bicycles, televisions, materials used for housing construction). These indices are normalized within each country, and we do no further normalization. The DHS has a similar 2-stage sampling design to the LSMS, and again we restrict clusters to those with multiple households and sufficient satellite imagery coverage. As in the LSMS, cluster locations in DHS are reported as the average latitude and longitude of cluster households, plus added noise to preserve the anonymity of survey respondents. As with consumption, we average the wealth index across households within each cluster.

We use data from the most recent DHS survey in five countries: Nigeria 2013, Tanzania 2010, Uganda 2011, Malawi 2010, and Rwanda 2010. Four out of five of these countries match those where LSMS data were available; we add the fifth (Rwanda) so that we can directly compare our results to those from a recent effort to estimate assets using cell phone records (*11*). Our final DHS dataset consists of 867 clusters in Nigeria, 455 clusters in Tanzania, 393 clusters in Uganda, 827 clusters in Malawi, and 492 clusters in Rwanda, for a total of 3034 clusters across the five countries. Each cluster in our sample contains between 11 and 45 households with a mean of 30.4 and a median of 27.

1.5 Nightlights data

Since the 1970s, the United States Air Force Defense Meteorological Satellite Program (DMSP) has deployed satellites equipped with Operational Linescan System sensors. Originally intended to monitor the global distribution of clouds and cloud-top temperatures, the nighttime satellite coverage has enabled the National Oceanic and Atmospheric Administration's National Geophysical Data Center (NOAA-NGDC) to isolate global human-generated lighting at some point between 8:30 and 10:00 pm local time every day.

The DMSP data processing entails removing observations distorted by cloud obstruction, moonlight, seasonally late sunsets, and auroral events. For every 30 arc-second cell, all remaining observations over the year are averaged and then converted to an integer "digital value" between 0 (no lighting) and 63 (representing top-coded luminosity) to produce a gridded satellite-year dataset. As of this writing, the DMSP public archive spans the years 1992 to 2013, with some years having two overlapping satellite

datasets (*20*). The accessibility, frequency, and unprecedented granularity of this data has sparked a recent literature using these "nightlights" as a proxy for economic activity and growth.

Beginning in 2014, the NOAA-NGDC has provided a separate nighttime lights dataset collected from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB). The VIIRS data is produced in a 15 arc-second geographic grid, twice the resolution of the existing DMSP product. However, these data have not yet been aggregated annually or been filtered to screen out temporal lights from aurora, fires, boats, and other transient sources, making it noisier and potentially biased. More importantly, VIIRS data do not exist for the survey years currently in our dataset. We decided not to use the higher-resolution VIIRS data for these reasons, but future processed versions of this dataset could be smoothly integrated into our model.

In our analysis, we use cluster-level coordinates provided by LSMS and DHS surveys to compare consumption and wealth predictions generated by our transfer learning approach to those derived from nightlights. To produce a nightlights estimate for a given location, we use the dataset corresponding to the year the survey was conducted, extract all digital values for cells within the 10 km by 10 km square centered on the provided coordinates, and assign the cluster the mean value.

1.6 Estimating consumption and assets

We evaluate ridge regression models in either 5- or 10-fold cross-validation for each country (as reported in the main text), for both consumption and for assets. Ridge regression is a linear regression model that enforces squared penalties on the size of the linear coefficients. In our case, since the dimension of our image features is large (d=4096), regularization helps to prevent overfitting to the relatively small training sets. The choice of regularization parameter for each fold was made in an inner cross-validation loop to preserve the integrity of the hold-out test data. The model r^2 reported is the average test r^2 across the cross-validation folds.

1.7 Randomization test

To confirm that the performance achieved by our model was not by chance, we conducted a "randomization inference" test in which we randomly shuffled the training labels so that each training example consisted of a randomly paired image feature vector and wealth label. Each trial evaluated a ridge regression model in 3-fold cross-validation, again choosing the regularization parameters in a nested cross-validation fashion, and after having reduced the dimensionality of the image features to 100 using PCA. For each country, we ran 1000 trials of this experiment and then plotted the resulting r^2 distribution and compared it to the predictive power of the corresponding true model trained on unshuffled data. All true models (i.e., models with correct cluster training labels) achieve statistically significant performance over randomized models at the 0.01 significance level.

1.8 Image features versus nightlights

We compared the performance of ridge regression models trained on image features extracted by the CNN against models trained only on nightlight values. While our primary goal is to compare the performance of our daytime image features against the standard use of nightlights is used in the literature—which is to

take an average or a sum of nightlight values over an area of interest, and to use that average as a proxy for or to predict some other economic outcome (*7, 10*)—we also evaluate models that make additional use of information in nightlights beyond area mean luminosity. See SM 2.2 for more details.

As before, model hyperparameters are chosen in a nested cross-validation fashion. Using 100-dimensional image features extracted by the trained CNN and reduced through PCA, we ran 100 trials of 10-fold cross-validation, using the same training and test sets to evaluate the image feature-based model and the nightlights model. We evaluate model performance using several metrics (r^2, root-mean-square error (RMSE)), and then compute the fraction of times where the image features outperform the nightlight features, the average margin of improvement, and the percent improvement.

We are often most interested in estimating well-being for the portion of the population that lives at very low levels of income. In many countries, the poorest are dispersed in rural areas where nightlights show little variation and thus have low predictive power (recall Fig. 1). To test whether the learned image features can provide us with information where the nightlights cannot, the feature comparison experiment was run on various subsets of the total available clusters in each country. For the LSMS countries, we first ran the experiment on clusters that had average consumption values below 2x the global poverty line, then repeated the experiment with thresholds of 3x, 4x, and 5x the poverty line, and then on all clusters. For the DHS countries, we also ran the experiment five times, on clusters below the 20th, 40th, 60th, and 80th percentiles, and finally on all clusters. For the pooled model, we ran comparisons restricting the sample to clusters below a given percentile in the consumption or asset distribution, starting with the 5th percentile and going up to the 100th percentile (full sample) in 5 percentile increments (Fig. 4a-b).

1.9 Out-of-country model generalization

Evaluating the out-of-country generalization of our models is important for understanding whether we can expect to make accurate predictions in countries where no survey training data is available. The distribution of landscape features could potentially be very different from country to country—this experiment gives us an idea of how robust our learned image features are to such variation.

We ran 10 trials of 10-fold cross-validation for each country, in which we trained the model using data from the base country, then evaluated the model on each of the other countries. The r^2 values reported are the average across the 100 total test folds. For the case when the pooled data was used for the base model, we ensured that the test data for each country was not included in the training data for the pooled model.

2 Additional results

2.1 Understanding performance differences for consumption versus assets

Our approach appears to more capably predict variation in assets than variation in consumption expenditures (Figs. 3, S3, S4). There are multiple potential explanations for this performance difference: *(a)* differences in the survey design, sample size, or survey quality between DHS (assets) and LSMS (consumption), *(b)* lower noise in the measured asset estimates compared to the measured consumption estimates, as the former is generated largely from survey responses that an enumerator can visually verify,

while the latter is generated from unverifiable responses to questions subject to substantial recall bias, and/or *(c)* the possibility that our image feature approach directly measures important assets used in construction of the asset score, such as variables related to household roofing materials.

To evaluate these possibilities, we use the available asset data in the most recent Uganda LSMS to construct our own asset index using principal components analysis (PCA) and following the methodology used by the DHS in the construction of their asset indices. To evaluate the possibility that image features were directly measuring roof characteristics, we create two versions of the asset index: one which uses all asset variables, and one which excludes any variables related to roofing materials or roof type. We then predict these asset indices using our image features, and compare performance on this task to performance on the consumption prediction task already reported. This experiment allows us to isolate explanations *(a)* and *(c)* from explanation *(b)*.

Using the "full" asset index (i.e., the one that includes roof materials), we achieve a cross-validated r^2=0.64 using our daytime image features as predictors, significantly higher than the r^2=0.41 achieved when estimating consumption using data from the same survey (Figure S5). Interestingly, predictive performance for assets does not decline when we reconstruct the asset index after excluding variables pertaining to roofing materials (r^2=0.65 on this task).

These experiments suggest that the difference in predictive performance between consumption and assets likely has more to do with lower noise in the survey-based asset measure compared to the survey-based consumption measure, and less to do with either underlying differences between DHS and LSMS surveys or an ability of the daytime imagery to directly measure important components of the asset index.

As further evidence refuting explanation *(c)* above, we use data released by DHS that describe results of the principal component analyses used by DHS to create their asset indices in each survey. Table S1 shows, for each of the five DHS surveys we use, the 10 variables in each survey that influence the asset score the most, ranked by the (absolute value) of each variable's component score in the PCA. Variables pertaining to roofing materials only rank in the top 5 variables in one out of five countries, and in the top 10 in 3 out of 5.

2.2 Additional nightlights results

While our primary goal is to compare the performance of our daytime image features against the standard way in which nightlights are used in the literature—which is to take an average of nighlight values over an area of interest and to use that average as a proxy for or to predict some other economic outcome—we also evaluate models that make additional use of information in nightlights. In particular, we evaluate models that used as additional regressors the median luminosity in the 10 km by 10 km region, as well as counts of pixels at different luminosity levels within the same region.

We find that using additional information beyond mean nightlights does indeed improve cross-validated performance of the nightlights models (Fig. S7, compare to Fig. 4a-b), although—as before—our transfer learning model still modestly outperforms nightlights for consumption throughout most of the consumption distribution, and substantially outperforms nightlights in predicting assets in the poorer part of the distribution.

While this experiment does not exhaustively explore the set of all possible ways in which nightlights could be used, it provides additional evidence that our image features contain information beyond what

Table S1: **Ranked importance of different assets in the DHS-constructed asset index in each country.** Data are derived from country files available on the DHS website (http://www.dhsprogram.com/topics/wealth-index/Index.cfm).

	Uganda		**Tanzania**		**Malawi**	
Rank	variable	score	variable	score	variable	score
1	Floor: dirt	-0.088	Floor: dirt	-0.087	Electricity	0.091
2	Electricity	0.082	Electricity	0.086	Roof: grass/thatch/mud	-0.090
3	Floor: cement	0.082	Electric lighting	0.086	Sofa set	0.090
4	Television	0.078	Television	0.081	Bed/mattress	0.089
5	Wall: cement	0.077	Floor: cement	0.081	Floor: dirt	-0.089
6	Wood for cooking	-0.076	Wood for cooking	-0.077	Television	0.089
7	Sofa set	0.071	Refrigerator	0.067	Roof: metal	0.088
8	Charcoal for cooking	0.067	Wall: cement	0.067	Bank account	0.082
9	Roof: grass/thatch/mud	-0.067	Roof: grass/thatch/mud	-0.066	Refrigerator	0.079
10	Cupboard	0.061	Charcoal for cooking	0.066	Mobile phone	0.073

	Nigeria		**Rwanda**	
Rank	variable	score	variable	score
1	Fan	0.085	Electricity	0.114
2	Television	0.081	Television	0.108
3	Flat iron	0.081	Floor: dirt	-0.107
4	Electricity	0.074	Charcoal for cooking	0.106
5	Bank account	0.073	Floor: cement	0.102
6	Floor: dirt	-0.071	Water piping into yard	0.091
7	Refrigerator	0.069	Domestic servant	0.083
8	Wall: cane, palm, trunks, dirt	-0.069	Computer	0.082
9	Wood for cooking	-0.067	Refrigerator	0.082
10	Kerosene for cooking	0.061	Mobile phone	0.070

nightlights provides.

2.3 Comparison to simpler methods of feature extraction from daytime imagery

Convolutional neural networks are now the de facto standard in the computer vision community to achieve state-of-the-art performance on many vision tasks. However, there also exist many other general methods for extracting features from images. One of the simplest methods for extracting image information is to take the average pixel value across each color channel. For the daytime satellite images that we worked with, this is simply the average RGB pixel value. Another method that makes use of color information summarizes the distribution of pixels in the image as a color histogram (i.e., by binning pixels into discrete bins based on their RGB values).

Yet another method takes the image as input and computes a histogram of oriented gradients (HOG). This HOG feature vector computes the gradients within the image by comparing the values of neighboring pixels, and then counts the occurrences of gradients of different orientations in localized areas of the image. The general idea is that image features of interest can be captured in the distribution of intensity gradients or edge directions. Finally, we also try directly reducing the dimensionality of the input satellite images through principal component analysis (PCA). PCA works by finding n orthogonal directions of maximum variation in the training data, and then representing all future data examples in the reduced n-dimensional space.

In Figure S8, we compare the performance of our transfer learning approach with these other approaches to extracting features from imagery. Due to the enormous complexity and variety of daytime satellite images, we find that the CNN model far outperforms all of the simpler approaches described above.

2.4 Comparison to approach using only survey data

We also study how our daytime image features perform against simpler ways that a policymaker might use available data to predict the current spatial distribution of some economic variable of interest. In particular, we study how our image features perform against an approach that uses available past surveys to predict current outcomes. This approach would only be feasible in the subset of countries with at least one survey (recall Figure 1), and can only be evaluated in the smaller subset of these countries with at least two surveys and with location coordinates available for both surveys.

Each of our five DHS countries had conducted at least one DHS survey prior to the survey used in our main analysis. Our survey-based approach takes each of the earlier surveys, creates a surface of asset scores for the country by spatially interpolating the cluster-level asset scores in that earlier survey (using inverse-distance weighting), and then uses this surface to predict the cluster-level asset scores observed in the most recent survey.

As seen in Table S2, interpolated data on cluster-level asset scores from past surveys are generally reasonably predictive of future asset scores, with predictive performance declining as the time between surveys increases. Nevertheless, our daytime image features perform roughly as well as data from the most recent prior survey (Uganda, Nigeria) or substantially better (Tanzania, Rwanda, Malawi). Given that our approach is substantially cheaper than undertaking an additional survey (ref (*5*) suggests that each DHS survey round costs roughly 1 million USD), and is applicable in any country, not just those who have undertaken a survey in the past, this experiment provides additional evidence that our transfer learning approach can provide useful additional information to policymakers.

Table S2: **Comparison of our model performance with predictive performance of interpolated earlier DHS surveys.** Second and third columns give the survey year and model performance of our image features. Last column gives the performance of the interpolated predictions based on earlier DHS surveys in the same country, with years of those surveys in parentheses.

Country	**Survey year**	**CNN r^2**	**Interpolated earlier survey r^2**
Uganda	2011	0.69	0.58 (2001), 0.70 (2006)
Tanzania	2010	0.57	0.40 (1999)
Nigeria	2013	0.68	0.43 (1990), 0.50 (2003), 0.70 (2008)
Rwanda	2010	0.75	0.72 (2005)
Malawi	2010	0.55	0.41 (2000), 0.45 (2004)

Figure S1: **Relationship between asset-based wealth index (from DHS) and nightlight intensity at the cluster level for five African countries.** Distribution of nationally-representative household-level wealth index scores shown beneath each panel in grey. Black lines are LOESS fits to the data with corresponding 95% confidence intervals in light blue.

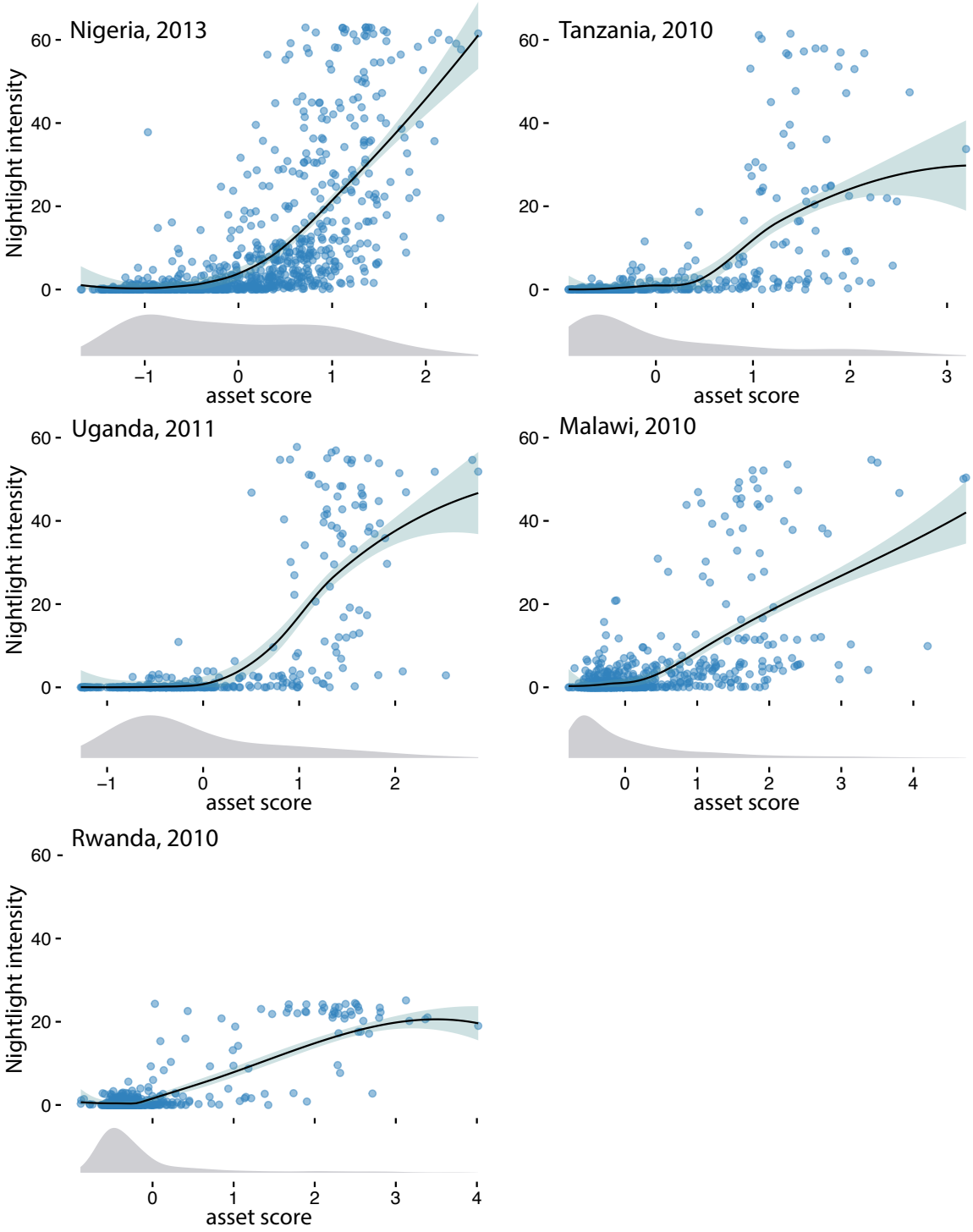

Figure S2: **Correlates of both poverty and nightlights that show variation around the poverty line and can potentially be remotely sensed. a.** Cluster-level relationship between consumption and either nightlights (blue line, as in Fig 1c-f) or distance from nearest population center with at least 100,000 people (green; reverse scale on right y-axis). The figure contains many urban clusters that have a "distance from nearest population center" equal to zero. Consumption data are from the 2011-12 Uganda LSMS survey and distance data is derived from the Global Rural-Urban Mapping Project (*27*). **b.** The same consumption and nightlights data, but with proportion of metal roofs derived from the LSMS survey (yellow). **c.** Nationally representative distribution of household-level consumption derived from the LSMS survey. Red vertical line in each panel denotes the international poverty line of \$1.90 person^{-1} day^{-1}.

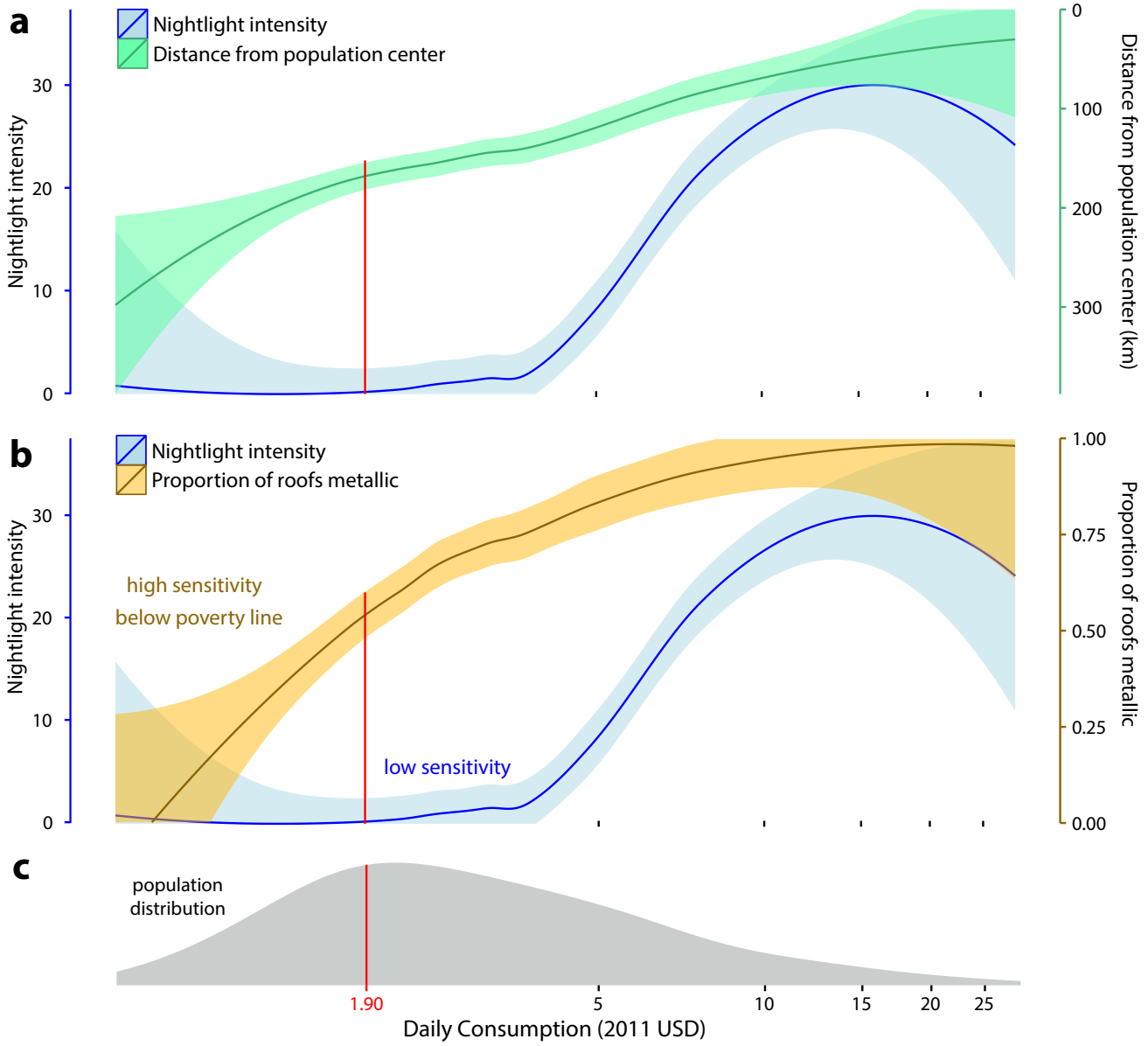

Figure S3: **Predicted cluster-level asset index from transfer learning approach (y-axis) compared to DHS-measured asset index (x-axis) for 5 countries.** Predictions and reported r^2 values in each panel are from 5-fold cross validation. Both axes shown in log-scale. Black line is the best fit line.

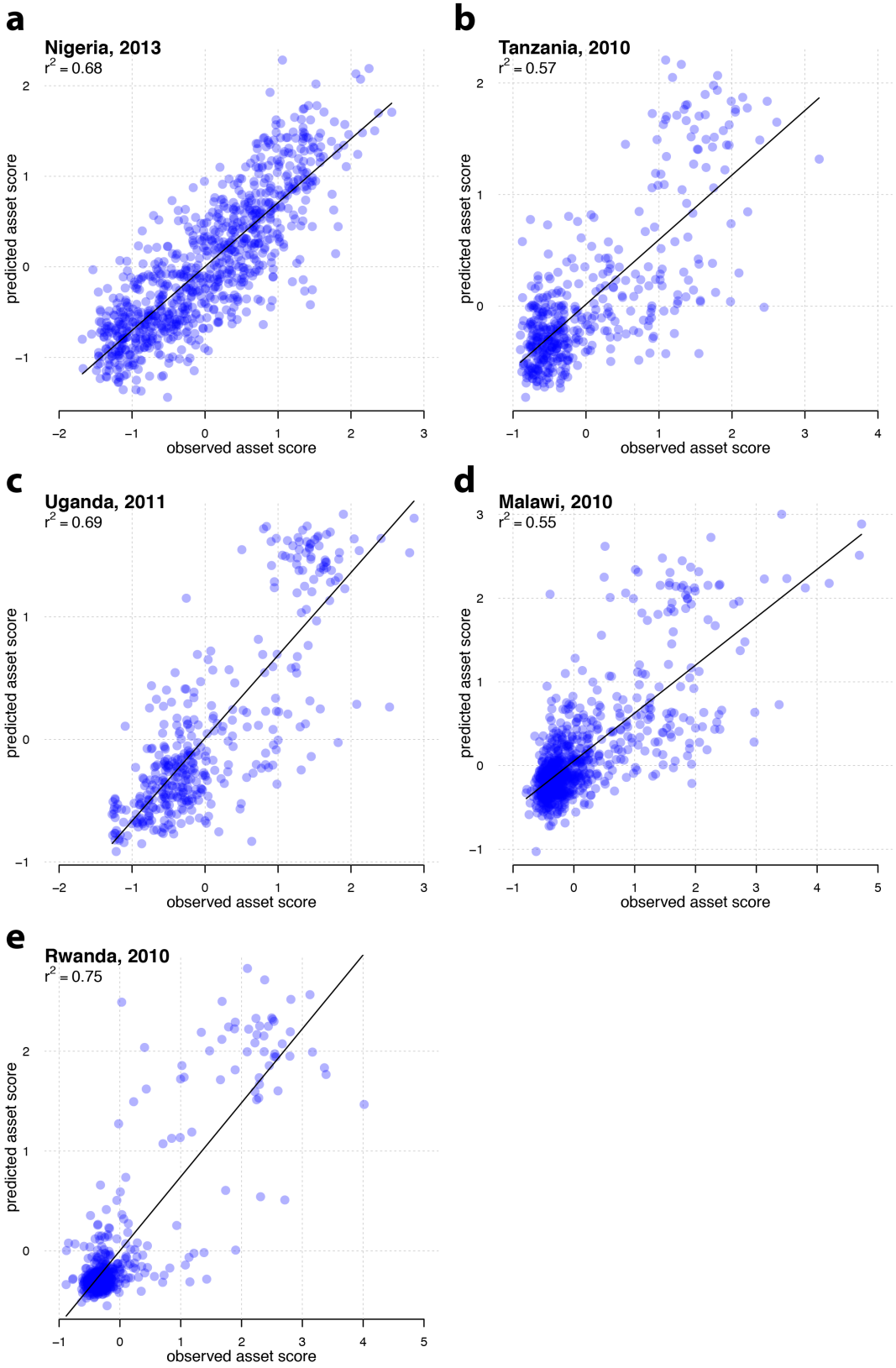

Figure S4: **Relationship between estimated and observed consumption (a) and assets (b)**, from a pooled model using data from all four LSMS countries (as in Figure 3) or all five DHS countries (as in Figure S4). Vertical red line in the left panel is the international poverty line ($1.90 person^{-1} day^{-1}). Both axes shown in log-scale for consumption.

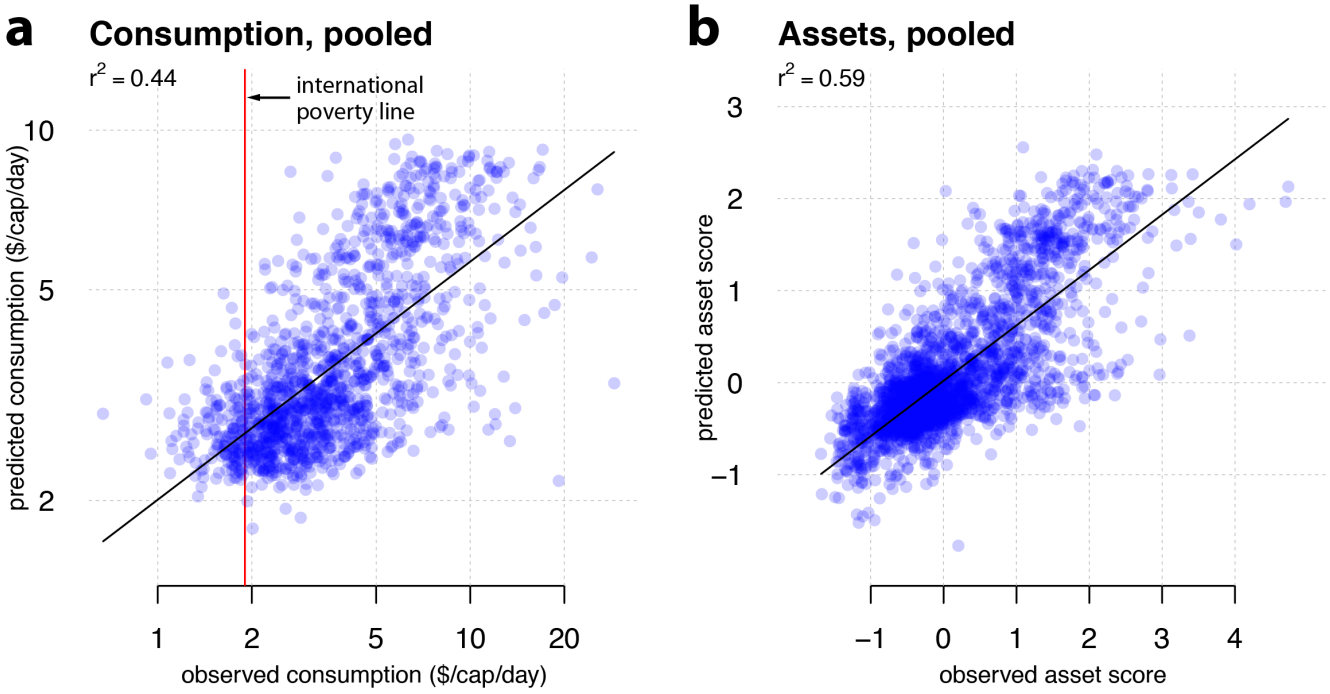

Figure S5: **Relationship between estimated and observed asset scores, Uganda LSMS**. Left panel uses an asset index that includes variables pertaining to roofing material, right panel omits these variables from asset index. Cross-validated r^2 are reported at top of each panel.

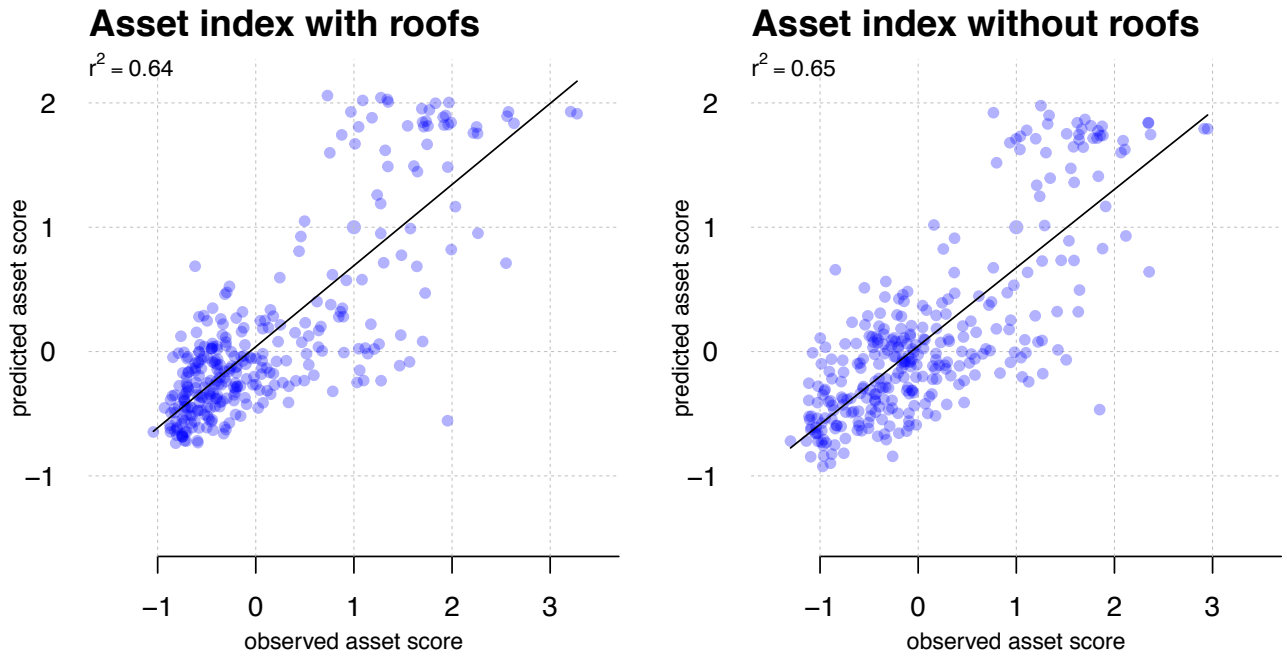

Figure S6: **Country-level performance of our model relative to nightlights.** For a given measure of model performance, each panel shows the percent of times our model outperformed nightlights across 100 trials of 10-fold cross-validation (x-axis), versus the average difference in the performance metric between our model and nightlights (y-axis). Trials were run independently for clusters falling below various multiples of the poverty line (for consumption) and various quintiles in the asset distribution. Rows show average r^2 improvement, percent r^2 improvement, average root mean squared error (RMSE) improvement, and percent RMSE improvement. Left column is consumption, right column is assets. Marker colors indicate countries, and size of marker indicates sample on which trial was run. Green shaded areas indicate models that outperform nightlights, while red shaded areas indicate models that underperform.

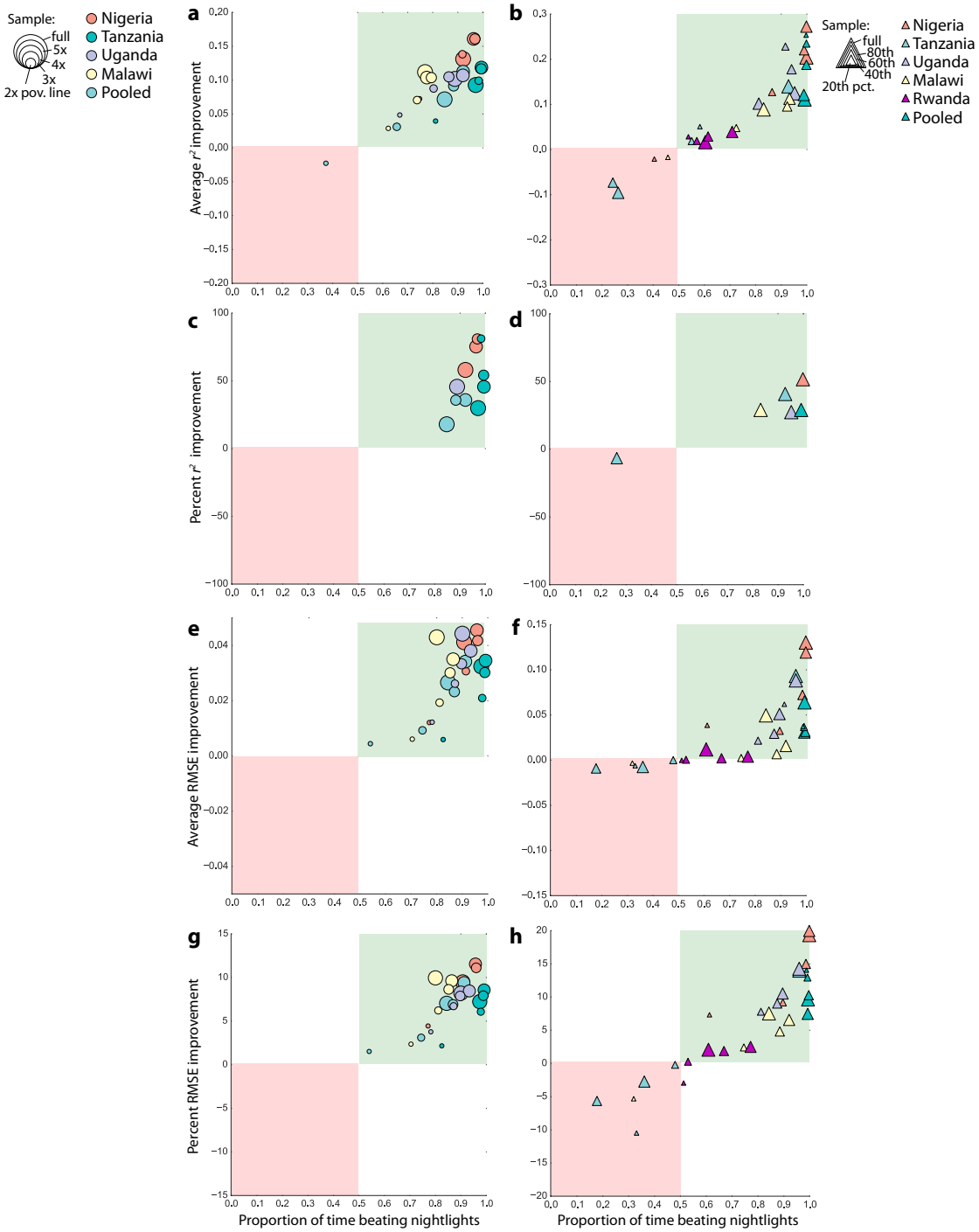

Figure S7: **Evaluation of model performance against nightlights**, using additional information in nightlights beyond mean luminosity (as in Fig 4a-b). See SM 2.1 for details. **a.** Performance of transfer learning model relative to nightlights for estimating consumption, using pooled observations across the 4 LSMS countries. Trials were run separately for increasing percentages of the available clusters (e.g., x-axis value of 40 indicates that all clusters below 40th percentile in consumption were included). Vertical red lines indicate various multiples of the international poverty line. Image features reduced to 100 dimensions using PCA. **b.** Same, but for assets.

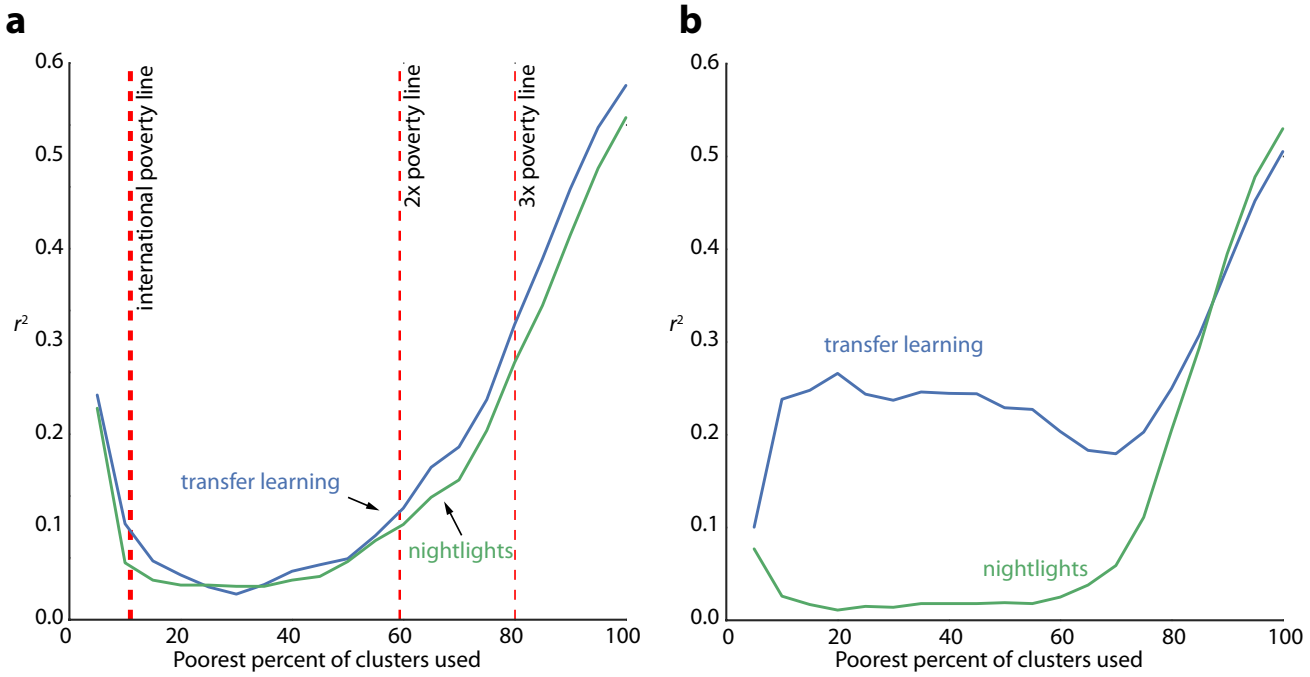

Figure S8: **Comparison of CNN and alternative feature extraction methods.** Bar heights represent cross-validated r^2 achieved using five different approaches to feature extraction from daytime satellite imagery. See SM 2.3 for details.