

# Measuring the Impacts of Poverty Alleviation Programs with Satellite Imagery and Deep Learning

Luna Yue Huang\*

This Version: October 6, 2020

## Abstract

Household surveys are expensive. In this project, I argue that housing, a strong correlate of wealth, can be accurately and cheaply measured with high-resolution satellite imagery and deep learning models, and can be used to reduce the costs of impact evaluations. In Western Kenya, I evaluate the GiveDirectly randomized controlled trial, a large unconditional cash transfer program, with satellite imagery, and observe statistically significant and economically sizeable increases in building footprint and roof quality, both useful proxy for housing quality. Using an Engel curve approach, I infer treatment effects on consumption and asset holdings, and obtain consistent results with extensive in-person surveys.

---

\*University of California, Berkeley; Doctoral Fellow at the Global Policy Lab. Contact: [yue\\_huang@berkeley.edu](mailto:yue_huang@berkeley.edu). I am extremely grateful to my advisors Marco Gonzalez-Navarro, Edward Miguel and Solomon Hsiang for their continued support and fantastic advising. I also benefited tremendously from suggestions and comments from Jeremy Magruder, Ben Faber, Marshall Burke, Joshua Blumenstock, Ethan Ligon, Elisabeth Sadoulet, Alain De Janvry, Aprajit Mahajan, and the participants in the AGU Fall Meeting 2019 (session: GC34C - Advances in Remote Sensing, Machine Learning, and Economics to Improve Risk Management and Evaluate Impacts in Socioenvironmental Systems), the UC Berkeley Trade Lunch, the UC Berkeley Development Workshop, the UC Berkeley Development Lunch, and the UC Berkeley Good Data Seminar. I thank Edward Miguel, Michael Walker, Dennis Egger, Johannes Haushofer, and Paul Niehaus, and the rest of the team who have worked on organizing the GiveDirectly randomized controlled trial, and collecting, cleaning and assembling the dataset for evaluation, for generously sharing the dataset with me and responding to my various inquiries. All errors are my own.

## 1 Introduction

Door-to-door household surveys have become an important basis for evidence based policy making in the field of international development. They generate crucial insights for the effectiveness of policies, but are expensive and time consuming to conduct. For example, a typical (median) impact evaluation funded by the 3ie (International Initiative for Impact Evaluation) costs around USD 334,000 and 3.4 years to complete. The costs of doing field surveys often limit the scope of impact evaluations to relatively small geographical areas or restricted samples. Political unrest, pandemics, and other unanticipated events often disrupt field surveys for long periods of time, leaving crucial data missing (1). Delays could potentially lead researchers to miss the optimal window for influencing policy (2).

Satellite imagery offers a promising avenue for timely and inexpensive data collection that may challenge the paradigm for impact evaluation in international development. They may be viable substitutes for field surveys in low-budget studies or real-time analysis; and could complement data collection efforts in the field by serving as pilots to derisk larger-scale projects or mid-line surveys in between more costly field surveys.

A recent literature using nighttime light intensity (hereafter “nightlights”) as a proxy for economic development has shown great promise of this approach on a relatively large geographic scale. Numerous studies illustrated the usefulness of satellite-based economic welfare measures, and how they opened up new research possibilities (3–22). However, the nightlights data have a few undesirable properties. First, it has poor sensitivity in less developed and rural regions with little to no electrification (23). For example, 99.73% of pixels are completely unlit in Madagascar, 99.47% in Mozambique, and this is representative of low-income countries (24). This presents challenges for development economists, when using night lights to study rural and/or less developed areas. The lack of sensitivity also raises the concern of false negative results and attenuated estimated effects when night lights data are used as the proxy for economic well-being. Second, night lights data have relatively low spatial granularity. Night lights also suffer from the “blooming” effect, where bright lights appear larger than they actually are, especially over water- and snow-covered areas. This induces strong spatial auto-correlation and introduces problems when one compares economic activity across contiguous small areas (25). Low spatial resolution prevents economists from exploiting fine spatial variations, such as those generated by most randomized controlled trials (RCTs), or those used in spatial discontinuity designs.

In this paper, I propose a method that combines insights from a classical economic concept, Engel curve, with high-resolution daytime satellite images and deep learning models to construct proxies for economic well-being that overcome these limitations. The deep learning model recognizes specific patterns from unstructured image data, and automatically extracts semantic information. I use a state-of-the-art deep learning model (Mask-RCNN) to conduct instance segmentation for building footprint on high-resolution satellite images, and directly and precisely measure housing

quality. An Engel curve describes the relationship between households' consumption of certain commodities and their income levels. There is a long tradition in microeconomics to leverage the stability of such relationships and cheaply measure income by observing just part of households' consumption basket: food, consumer durables, etc. Building on prior work (26–28), I infer changes in overall economic welfare from observed changes in housing quality. The Engel curve approach is appropriate for housing, as housing quality usually increases as income increases, and it accounts for a large proportion, around 10–20%, of people's overall spending.

As a proof of concept, I evaluate a randomized controlled trial in western Kenya, where households are randomly assigned large unconditional cash transfers from GiveDirectly. I demonstrate the feasibility of conducting impact evaluation with two proposed housing quality measures: building footprint, and roof quality. In many developing country contexts, metal roofs, which are more durable and higher-quality than thatched roofs, are often a good proxy of wealth.

Estimation with night lights yields a null result, missing large positive impacts of the treatment on recipients. This may be because of low demand for electrification, which is corroborated by field surveys, the fact that nightlights are not sensitive enough to detect variations in wealth or expenditure in less developed regions, or the fact that nightlights have a lower spatial resolution, potentially reducing statistical power.

On the other hand, I observe statistically significant and economically sizable effects on housing quality, on both the intensive margin (higher quality roofs) and the extensive margin (larger building footprint). In cases where the intervention does not change the Engel curve, I am able to infer overall treatment effects on household consumption and assets, as measured in household surveys in the field, from observed increases in housing consumption. I obtain consistent results with extensive in-person surveys, although the estimates are less precise.

As an additional check in a different cultural context, I validate satellite-derived housing measures against the 2010 Population and Housing Census records of rural localities (villages) in Mexico. I show that, reassuringly, the predicted number of buildings from daytime satellite images is highly correlated with population count in the census, and does not display diminished sensitivity in smaller and poorer rural localities. Average building size is correlated with asset scores reported in the census, indicating that these measures are useful proxies of wealth.

This paper builds on prior empirical work, which has started exploring the use of satellite-derived housing quality measures such as roof reflectance (29), roof color and alignment of buildings (30), in evaluating specific policies. Scholars in other fields have also pursued the idea of using building footprint to study wealth, and in particular, using the distribution of building sizes within a settlement to study wealth inequality and social stratification (31–33). I develop the methods more formally, and relate these housing quality measures to overall economic welfare with an Engel curve approach.

The methods proposed here are related to an emerging literature in poverty mapping. Echoing

an earlier literature in development economics (34), several recent studies show that high-resolution satellite images can be used to create spatially granular poverty maps (23, 35–40). These poverty maps are much cheaper to compile than conventional methods, and are useful for precisely targeting foreign aid or domestic assistance programs. In this paper, I propose to adjust these methods such that they are more appropriate for impact evaluation. These methods obtain poverty estimates by first training the machine learning model to predict living standard measures from satellite images, using census records, LSMS/DHS household expenditure surveys or researcher-collected survey data as the ground truth labels. To use these estimates as proxies of overall economic well-being for impact evaluation, one has to implicitly make the assumption that the intervention does not change the underlying relationship between extracted features on the images (e.g., infrastructure, agricultural land, etc.) and overall economic well-being. While this assumption is sometimes appropriate, it can be challenging to satisfy in certain settings — for example, if satellite-derived poverty estimates are used to evaluate infrastructure improvements, or education interventions, these interventions not only change recipients’ living standards, but also the composition of their consumption basket. In this paper, I propose to disentangle (1) the prediction exercise from pixel values on raw satellite images to semantically meaningful observations of housing and housing quality, a mechanical relationship that is unlikely to be changed by interventions, and (2) the prediction exercise from observations on housing consumption to overall economic well-being, a relationship that is easier to theorize about and explicitly test. This frames the first task as a traditional computer vision task that could leverage extensive research and recent development by the computer science community. This also makes the second assumption more explicit and testable.

## 2 Methods

Figure 1 demonstrates the overall data pipeline in this paper.

As a proof of concept, I evaluate a randomized controlled trial in rural Western Kenya. The experiment was conducted with GiveDirectly, a US charity, and randomized cash transfers to households in about half of the 653 villages in the experimental sample. From 2014 to 2016, around \$1000, which is equivalent to about 75% of the annual household expenditure, was distributed to households in lump sum with no conditions (see Appendix Section B.1.1 and the original paper (41) for more details on the design and implementation of the trial). In Panel a, I illustrate the spatial variations in treatment intensity that were generated by the trial. Each point is the GPS coordinate of a trial participant, and they are randomized into the treatment group or the control group.

I overlay a regular grid with the GPS coordinates of trial participants (Panel b). I calculate how many participants received the cash transfer, as well as the housing quality in each grid cell (Panel c). The housing quality metrics are computed from deep learning model predictions (Panel e) on corresponding high-resolution satellite imagery (Panel d), as described in more detail below.

Using grid cells as the unit of observation, I estimate the average treatment effect of the GiveDirectly cash transfer on remotely sensed housing quality metrics. Using an Engel curve, which describes the relationship between an aggregate economic well-being measure and housing quality, I infer the average treatment effects on the aggregate measure (Panel f).

## 2.1 Satellite Imagery and Deep Learning

**Satellite Imagery** The input high-resolution satellite images are from Google Static Maps (42). These images have been pre-processed (e.g., to remove clouds) and geo-referenced and they come from a variety of sources such as Maxar (formerly DigitalGlobe) and Airbus. Unlike other remote sensing data sets, these images only contain the RGB (red, green, blue) bands, and no information in other wavelengths.

The input satellite images are at a spatial resolution of about 30 cm per pixel (on equator).

Google Static Maps uses satellite images from different commercial providers and different time periods, and seamlessly mosaic images together. Most images are taken fairly recently. However, no exact timestamps are available for the images.

To evaluate the GiveDirectly randomized controlled trial in Kenya, I download a set of satellite images covering all of the Siaya county in Western Kenya, where the study took place. The images were retrieved from the Google Static Maps API between Feb 19 and Feb 21, 2020. Given that the GiveDirectly cash transfers were given out from mid 2014 to early 2017, it is safe to assume that most of these satellite images are taken after the intervention. This is consistent with the fact that I observe strong treatment effects.

**Mask R-CNN** I leverage a state-of-the-art instance segmentation model in computer vision to identify buildings and outline their boundaries. Specifically, I used a model named Mask R-CNN (43). Intuitively, the model first examines the satellite images and extracts feature at different scales; it then identifies a large number of regions of interest; finally, it predicts whether the regions of interest contain an object and, if so, classifies the object and generates pixel masks for each object. Appendix Section B.2 describes this step in more detail.

One of the main challenges in training building footprint instance segmentation models in developing countries is that training data are sparse. Most mainstream data sets on building footprint, such as SpaceNet and Inria Aerial, consist of images in metropolitan cities in developed countries (e.g., Chicago, Las Vegas, Paris and Shanghai). These urban landscapes are dramatically different from most rural scenes in my analysis sample. Additionally, both SpaceNet and Inria Aerial provide semantic segmentation labels, and do not annotate buildings that border each other separately.

To overcome this challenge, I train the Mask R-CNN model with a multi-step process, utilizing both publicly available data sets and data sets created specifically for this paper. First, I pre-train the model on COCO (Common Objects in Contexts), a large natural image dataset (44). Second,

I pre-train the model on Open AI Tanzania, a collection of overhead images taken by consumer drones in Zanzibar, Tanzania, that are annotated with building footprints (45). These images are more representative of the rural or peri-urban scenes in a developing country context, in terms of the distribution of the density, sizes and heights of the buildings. Third, I create in-sample annotations by randomly sampling from the set of images that predictions will be made on, and annotating them (see Appendix Section B.1.3 for a more detailed description). This ensures that training images and inference images belong to the same data distribution. Because the set of in-sample annotations is small, I rely on transfer learning (i.e., pre-training on larger data sets) and extensive data augmentation to prevent overfitting.

**Model Performance** Figure 1 shows an example high-resolution satellite image (Panel d), and extracted building footprints, as well as an average roof color profile (Panel e), which is later used for computing housing quality measures. Figure A.1 show 10 more examples of my prediction results, which are randomly sampled from the entire archive of satellite images (in the study area in rural Kenya). Aside from qualitative results, I also provide quantitative performance evaluation of the Mask R-CNN model. The performance metrics reported in this section is representative of that on the entire GiveDirectly analysis sample, because the evaluation is conducted on an annotated random subset of the GiveDirectly images (see Appendix Section B.1.3 for details).

Figure A.2 shows the precision-recall curve on the in-sample annotations in Siaya, Kenya, computed with three-fold cross validation. The F1 score of the model is 0.79, higher than the performance observed in some modern satellite imagery benchmarks (see Appendix Section B.2.3 for more details).

**Extraction of Housing Quality Metrics** I extract two measures of housing quality: building footprint, and roof quality.

First, I compute the area covered by each predicted building polygon.

Second, I extract an “average” roof color for every building, as shown in Figure 1 (Panel e). Roof color is often a good proxy for roof materials, and has been used by prior work to measure housing quality (29).

To improve the interpretability of roof colors, I cluster these roof color vectors into 8 groups by running the K-means clustering algorithm. Figure A.3 show the 8 color clusters formed, and I group them further into tin, thatched, and painted roof materials. The color on the left represents the “average” color, and the colors on the right represent a random sample of all the colors belonging to this cluster. The number of squares on the right is proportional to the number of observations belonging to the clusters.

## 2.2 Average Treatment Effect Estimation

To conduct the estimation, I lay out a regular grid with a spatial resolution of  $0.001^\circ$  or 0.11 kilometers (with longitudes ranging from  $34.03^\circ$  E to  $34.46^\circ$  E and latitudes ranging from  $0.06^\circ$  S to  $0.32^\circ$  N). I create rasters of treatment intensity, and outcomes of interest, as shown in Figure 2. For better display, Figure 2 shows a slightly higher spatial resolution of  $0.005^\circ$  or 0.55 kilometers.

I compute treatment intensity based on data acquired from the original authors of the trial (41), and the data acquired are documented in more detail in Appendix Section B.1.1. The authors conducted a baseline census, and recorded the GPS coordinates of all the households that reside in the 653 villages within the study area. In each village, about 1/3 of the households met the eligibility criteria for the transfer. The trial uses a two-tier randomization technique to generate random “saturation” of the treatment in different areas. Each saturation group, which includes on average approximately 10 villages, was randomly assigned a high or low treatment saturation level. In high saturation groups, two-thirds of the villages are assigned to treatment, and the rest to control. In low saturation groups, one-third of the villages are assigned to treatment, and the rest to control. Households in the control villages never received the cash transfers; households in the treatment villages received the transfer if they met the eligibility criteria set by GiveDirectly. To construct the treatment intensity map, I count how many households received the transfer in each pixel, masking out pixels where no eligible households live. Figure 2 (Panel a) shows the number of eligible households that received the transfer.

I aggregate satellite-derived housing quality measures to the pixel level. In each grid, building polygons are assigned to pixels based on polygon centroids. I compute the total area covered by building as (overall) building footprint, and the total area covered by tin roofs as tin-roof areas<sup>1</sup>.

The econometric specification is as follows

$$y_i = \sum_{k \in K} \tau_k \mathbf{1}\{x_i = k\} + \sum_{m \in M} \beta_m \mathbf{1}\{e_i = m\} + \epsilon_i \quad (1)$$

where each observation  $i$  represents a pixel covering a area of about 0.012 square kilometers;  $x_i$  denotes the number of recipient households per pixel (equivalent to the amount of cash infusion, as each household received a transfer of roughly \$1000), with  $k \in K = \{1, 2, 2+\}$ ;  $e_i$  denotes the number of eligible households per pixel with  $m \in M = \{0, 1, 2, 3, \dots\}$ ; and  $y_i$  denotes satellite-derived housing quality outcomes, such as building footprint or roof reflectance. In order to utilize the variations in treatment intensity generated by the randomization, and account for effects of pre-existing population density or living standards, I flexibly control for the number of eligible

---

<sup>1</sup>I refrain from computing the proportion of tin roofs, because the expected treatment effects on that variable is ambiguous — households may upgrade the roof of their primary residence, which increases the proportion of area covered by tin roofs, or they may build new auxiliary structures such as a shed or kitchen (typically with thatched roofs), which decreases the proportion of area covered by tin roofs.

households per pixel. Pixels with no eligible households are dropped. Because the pixels are fairly small, the number of observations for  $k > 2$  is small. As such, I bin the number of recipient households into four bins  $\{0, 1, 2, 2+\}$ , to preserve statistical power. Standard errors are calculated à la Conley, with a uniform kernel and a 3km cutoff (46–49). To fully illustrate that this specification is identifying the causal effect of the cash transfer, I run 100 placebo tests that re-simulate the original randomization scheme. Specifically, in each simulation, I take the 68 saturation groups, randomly assign half to high saturation group and half to low saturation group; then I randomly assign villages to treatment or control groups based on these simulated saturation statuses. I estimate the placebo treatment effects using these simulated placebo treatment statuses.

To compute a single pooled effect, I assume linear treatment effects—every transfer of \$1000 has an effect of the same magnitude, regardless of treatment intensity in that geographical area. As such, the econometric specification is as follows

$$y_i = \tau x_i + \sum_{m \in M} \beta_m \mathbf{1}\{e_i = m\} + \epsilon_i \quad (2)$$

where  $\tau$  is the pooled (average) effect, and all else remains the same as above. These pooled estimates and their confidence intervals are reported in the panel titles in Figure 3.

To systematically compare the results with nightlights, I conduct the same estimation with the nightlight (VIIRS-DNB data product) raster in 2019. In this case, the average treatment effect is estimated using the native nightlight raster grid, with a resolution of 15 arc seconds, or roughly 463 meters.

### 2.3 Engel Curve

**Theoretical Framework** In this section, I lay out a model that is based on canonical economic models of Engel curves (26, 27) and simplified to fit the context of this paper. Let some household  $h$  with the welfare measure  $W_h$  (be it expenditure, assets, or other measures of economic well-being) consume  $Q_{hp}$  quantities of normal good  $p$ . Formulate the relationship between  $Q_{hp}$  and  $W_h$  as

$$Q_{hp} = F_p(W_h) + \epsilon_{hp} \quad (3)$$

where  $F_p(\cdot)$  is strictly increasing, meaning that households consume more of product  $p$  as their welfare measures increase. Simplifying this with a linearity assumption

$$Q_{hp} = \alpha_p + \beta_p W_h + \epsilon_{hp} \quad (4)$$

where  $\alpha_p$  is a constant and  $\beta_p$  is the coefficient of a linear Engel curve.

Suppose that one is interested in studying the effect of a plausibly exogenous treatment  $Z$  on

$W$  (denoted  $\hat{\tau}_W$ ), but one can only inexpensively observe its effect on  $Q_p$  (denoted  $\hat{\tau}_{Q_p}$ ). Assuming that  $F_p(\cdot)$  does not depend on  $Z$ , then one could re-scale the treatment effect estimate on the observable product  $p$  based on their knowledge of the Engel curve

$$\hat{\tau}_W = \hat{\tau}_{Q_p} / \hat{\beta}_p \quad (5)$$

This formulation is adapted from prior work that uses survey measures to infer real income growth (26). The key assumption in this exercise is that there exists a stable Engle curve. This is sometimes termed the conditional independence assumption (27). The intuition is that, as households' income level changes, potentially as a response to interventions, their consumption basket composition is stable, conditional on their income level. This assumption will be violated if certain interventions change households' behavior in how they spend their money, for example, food voucher programs may increase households' overall food consumption, making food-based Engel curve imputations unreliable.

Using a formula for propagation of error (or the multivariate Delta method), one can derive the standard error for  $\hat{\tau}_W$  as follows.

$$\left( \frac{\hat{\sigma}(\hat{\tau}_W)}{\hat{\tau}_W} \right)^2 = \left( \frac{\hat{\sigma}(\hat{\tau}_{Q_p})}{\hat{\tau}_{Q_p}} \right)^2 + \left( \frac{\hat{\sigma}(\hat{\beta}_p)}{\hat{\beta}_p} \right)^2 \quad (6)$$

This is in spirit similar to the exercise in prior work (26), but additionally accounts for the precision of the beta coefficient in the Engel curve. The precision of the final estimate depends on two factors: the precision of the treatment effects estimated from satellite imagery, and the precision of the Engel curve beta coefficient. As such, it is necessarily less precise than estimates based on field surveys if field survey covers the entire sample, because it incorporates noises arising from the Engle curve estimation exercise. However, it may be more precise if the field surveys only sampled a small fraction of the entire sample, because the satellite-based treatment effects estimation will be based on a larger sample.

In this exercise, I obtained the Engle curve estimates from the actual experimental data, because the GiveDirectly team conducted extensive in-person surveys to measure the effects of the intervention on various economic indicators. One could alternatively conduct these Engle curve estimations from a different sample, such as from a representative sample in LSMS/DHS surveys.

**Empirical Setup** From the endline survey data, I obtain high-quality measures of economic well-being, including total consumption and assets, constructed from a comprehensive consumption and expenditure survey (41). The survey analysis sample that I use contains 7,987 households in 653 villages, of which 4,578 households are eligible for the transfer. Household consumption expenditure is the annualized sum of total food consumption in the last 7 days, frequent purchases in the last month, and infrequent purchases over the last 12 months. Asset measures include

livestock, transportation (bicycles, motorcycles, and cars), electronics, farm tools, furniture, other home goods, and lending/borrowing from formal/informal sources. Asset measures also include housing value, measured as the respondent's self-reported cost to build a home like theirs, and land value, measured as landholdings multiplied by the self-reported per-acre cost of land of similar quality in their village. I compute consumption or assets per capita by dividing the household consumption or asset values by the number of household members (including children).

Before estimating treatment effects of the intervention, or Engel curves, I need to perform heuristic matching between buildings (as identified in satellite imagery) and household GPS coordinates. GPS coordinates are often noisy, and it can be challenging to assign structures to households in a sensible manner.

To do this, I merge both the household GPS coordinates in the endline survey and the coordinates of detected structures on satellite imagery to the household GPS coordinates in a baseline census. I acquired the baseline census data from the original authors, which covered all the household living in the sample area, and constructed a Voronoi diagram with all the GPS coordinates. Each structure identified in the satellite imagery, or each survey with a GPS coordinate, is assigned to its nearest census GPS coordinate. If the distance between the two coordinates is larger than 50m, the observation is discarded. In cases of multiple surveys being assigned to the same census coordinate, I keep the closest survey. This matching scheme ensures that each structure or survey is matched to at most one household, eliminating concerns for double counting.

In order to facilitate direct comparison with observed treatment effects on aggregate economic well-being measures, I re-estimate the treatment effects, using each household as the unit of observation. I estimate  $\hat{\tau}_{Q_p}$ , the treatment effects on housing quality, via a simple linear regression, using the eligible sample, and with the binary treatment dummy being the only right hand side variable<sup>2</sup>. I estimate the treatment effects on the aggregate economic well-being measure similarly. All the variables used are converted to per capita terms, so the household size variable is only available from the endline surveys. Therefore, the samples used for generating these estimates are restricted to households that can be linked to an endline survey, and the sample size can be much smaller than the preferred specification (with grid cells / pixels as the unit of observation).

I estimate the Engel curve using the full sample, including both eligible and ineligible households, and via a simple linear regression of a housing quality metric on an aggregate economic well-being measure<sup>3</sup>. While the Engel curve can theoretically be of any functional form, in practice, a linearity assumption is appropriate and helps maximize statistical power when both the X and Y variables are appropriately transformed. Consumption or assets per capita is shown in log scale, as is standard in the economic literature. Since these variables have been winsorized, no negative or zero values exist. Building footprint and tin-roof area per capita are also shown in log scale<sup>4</sup>.

---

<sup>2</sup>I do not perform any clustering adjustments on the standard errors.

<sup>3</sup>I do not perform any clustering adjustments on the standard errors.

<sup>4</sup>More precisely, I convert the variables with the function  $\log(x + 1)$  and discard the zero values.

In the final step, I compute the “estimated” treatment effects on aggregate economic well-being measures, based on the prior estimates and with Equation 5 and 6. I then compare those against the observed treatment effects, as measured in the endline surveys.

### 3 Results

**Spatial Distribution of Treatment Intensity and Housing Quality** Figure 2 shows the spatial distribution of treatment intensity (Panel a) in the study area. The white lines outline the GiveDirectly study area in Western Kenya. Within the area, the experiment follows a two-level randomization design to generate random variation in whether a village enters the treatment group, and also how many of its neighboring villages enter the treatment group. Under this design, the experiment generates many clusters of high treatment intensity areas, and low treatment intensity areas. The white pixels are geographical areas where no eligible participants live, which are either town centers that are excluded from this experiment, or uninhabited areas.

Panel b and c of Figure 2 shows the spatial distribution of two housing quality metrics: building footprint, and tin-roof area (area covered by high-quality tin roofs). Both of the maps are consistent with the geography of the study area, with red spots corresponding to towns, and the blue area on the bottom left corresponding to lakes and uninhabited natural areas.

**Average Treatment Effect Estimation** From these maps, I estimate the average treatment effects of the GiveDirectly cash transfer on multiple remotely sensed variables, and show them in Figure 3. Following the empirical specification in Equation 1 and 2, I plot the changes in housing quality metrics as cash inflow intensity increases (shown in red). As an additional check, I randomly simulate the treatment assignment process 100 times, and estimate the effects of these “placebo treatments” (shown in grey). Reassuringly, the placebo treatment effects are centered around zero, and are consistent with the main specification.

The average treatment effects on normalized nightlight values is small in magnitude (0.002), and not statistically significant. This may be because of low demand for electrification (50), the fact that nightlights are not sensitive enough to detect variations in wealth or expenditure in less developed regions, or the fact that nightlights have a lower spatial resolution, potentially reducing statistical power.

However, the average treatment effects on both building footprint and tin-roof areas are economically sizable, and statistically significant (Panel b and c). This shows that housing quality metrics for recipient households are increasing both in the intensive margin (quality improvements are made on existing houses), and the extensive margin (new structures — either primary residence or auxiliary structures, such as kitchen sheds — are built). This is in line with results based on extensive field surveys, finding that on average, overall assets (excluding housing and land values)

increased by \$180, housing values increased by \$378, as a result of the cash transfer<sup>5</sup>.

**Engel Curve** Finally, I illustrate that these observed housing quality improvements can be used to infer treatment effects on overall household asset holdings (Figure 4) and household consumption (Figure A.4). In Figure 4, the grey line shows the relationship between overall asset holdings (as measured in endline surveys) and normalized nightlight values (Panel a), building footprint per capita (Panel b) or tin-roof area per capita (Panel c). All the variables except nightlight values are shown in logarithmic scale, to best satisfy the linearity assumption of the Engel curve. Interestingly, the normalized nightlight values are not clearly positively associated with wealth, potentially indicating that this variable is not sufficiently sensitive or granular to pick up precise signals on wealth. On the other hand, remotely sensed variables constructed from daytime high-resolution satellite images show a clear monotonic relationship with wealth.

As discussed above, a key assumption when using this proposed method is that the treatment per se does not change the Engel curve. This seems plausible in the context of this experiment. The eligibility criterion for the transfer is that households live in thatched-roof houses<sup>6</sup>. It seems plausible that households may have psychologically tied the transfer to roof upgrading, because of the eligibility criteria. The lump sum nature of the transfer may also have led recipient households to spend more on lumpy purchases than they normally would. To test this assumption, I plot two separate Engel curves for household who have, or have not received the cash transfer. The two Engel curves do not appear to be different, indicating that the assumption may be valid.

As such, I compute the “estimated” treatment effects on overall asset holdings based on Equation 5 and 6. For each remotely sensed proxy, I compute one separate estimate. The estimate based on normalized nightlight values is both wrong and imprecise, potentially due to both null effects in the average treatment effect estimation, and the undesirable properties of its Engel curve. The estimate based on both of the housing quality metrics is broadly consistent with, albeit less precise than, the observed result in the survey.

I present the same set of analysis on household consumption in Figure A.4, and obtain qualitatively similar conclusions.

**Additional Evidence from Rural Mexico** To check the robustness of these measures across cultural contexts, I perform a similar analysis in rural Mexico, with the 2010 Population and Housing Census (Figure A.5). I compute two remotely sensed variables, the number of houses and the average size of houses, and check whether they are correlated with corresponding measures in

---

<sup>5</sup>The original paper found large spillover effects to non-recipients on consumption but not on assets, and I do not examine the spillover effects in the present project.

<sup>6</sup>Recipients in the GiveDirectly trial may be expecting subsequent transfers and thus may strategically avoid upgrading their roofs from thatched roofs to metal roofs, so that they would be eligible for potential future transfers. However, there is very little evidence suggesting that that was the case, and the researchers were clear that the intervention was not recurrent and that no future interventions were being planned.

the census (see more technical details in Appendix Section B.1.1 and B.2).

Remotely sensed variables based on daytime high-resolution imagery (shown in the right panels) perform consistently better than nightlights (shown in the left panels). Population size in a rural village, as reported in the 2010 census, is highly correlated with the number of houses in that village — with a Pearson correlation coefficient of 0.82 — but only modestly correlated with nightlight values (Panel a and b). Nightlight is less sensitive in smaller, less populated villages, which is consistent with prior findings (23). The asset scores of each village, measuring overall wealth, displays a stronger correlation with average size of houses than with nightlight values (Panel c and d). This is not driven entirely by housing consumption being a component in overall wealth. In fact, durable asset holdings, which are not directly visible on satellite images, are also equally strongly correlated with the average size of houses (Panel e and f). This is in line with the Engel curve intuition — a component of a household’s consumption basket can reveal information about their overall economic well-being.

## 4 Discussion

In this paper, I show that for lower income, and more rural contexts, high-resolution daytime satellite imagery can be a better proxy for economic well-being than nightlights. However, nightlights may demonstrate good sensitivity in middle-income countries, or in highly urbanized areas in low-income countries. There may also be instances where the two sources of data will generate complementary signals that can be fused to form better predictions or proxy measures.

However, a fundamental limitation of using any remotely sensed variables for impact evaluation is that, some outcomes are never observed in overhead imagery: education, health, ownership of land, etc. While I show that satellite-derived outcomes are correlated with some outcomes that cannot be directly observed, such as consumer durables inside houses, effects of certain types of interventions (e.g., teacher training) are bound to be difficult to observe from space.

We also cannot observe the movement of people in response to positive or negative economic shocks from satellite imagery, which is often essential for interpreting the effects of certain policies or interventions. While migration rates are low in the GiveDirectly study areas, and this doesn’t constitute a major concern in the experiment studied in this paper, it could present major challenges if used to study other types of interventions that naturally impacts mobility.

Remotely sensed variables can be context specific, and less directly interpretable than explicit survey measures. For example, the assumption that most buildings are one storey high is appropriate for most rural areas, but may quickly become questionable as we move to an urbanized context. The fact that wealth is generally associated with bigger houses is also mostly appropriate, but can fail in highly dense urban areas, where land supply is restricted and housing prices are high. The use of imputed types of roof materials to quantify roof quality, and thus housing quality, is valid in low-income contexts, but not necessarily in middle- or high-income contexts, where most roofs are

made of durable materials, and roof colors are more of a reflection of personal taste and culture, rather than wealth or social status.

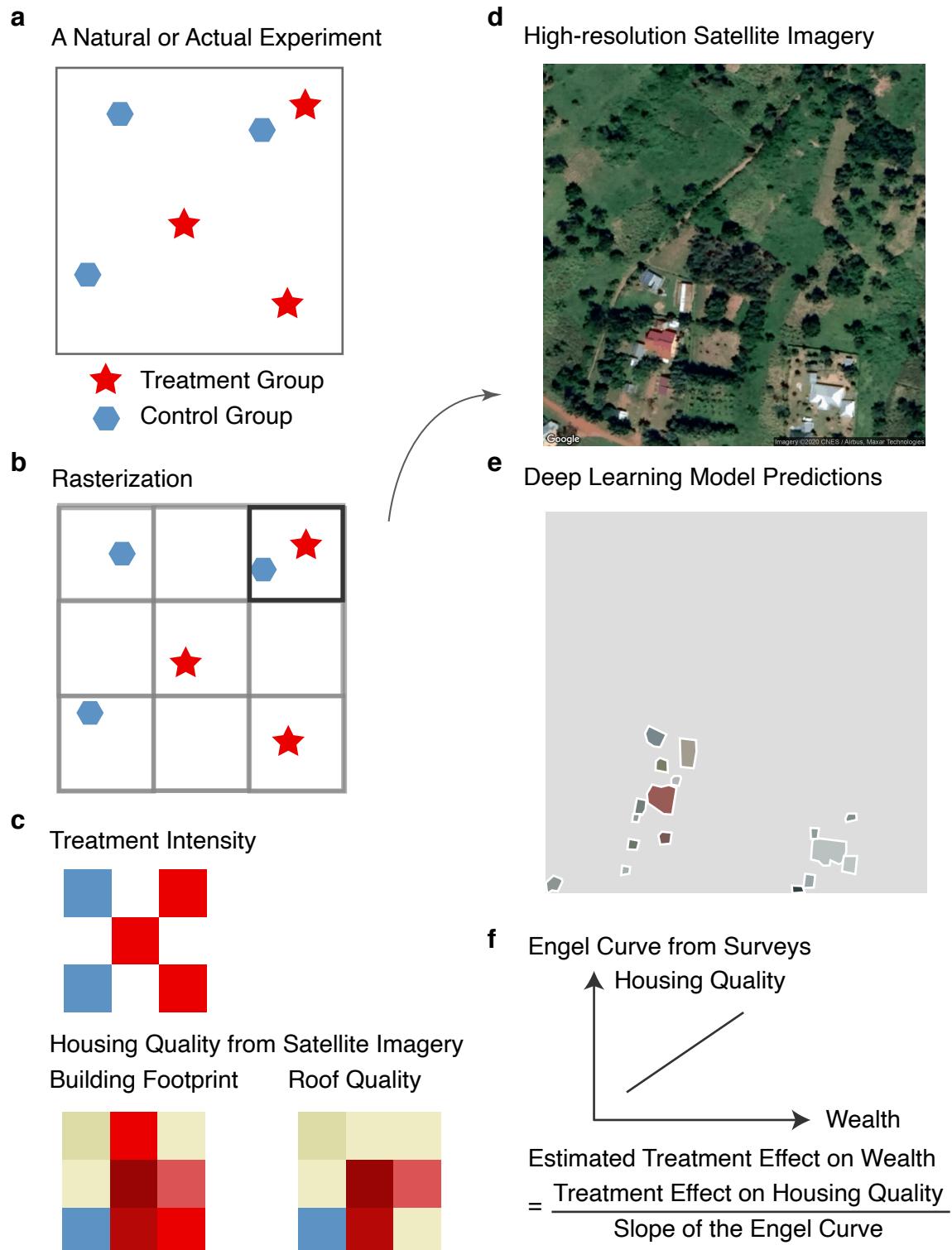


Figure 1: Data pipeline schematic.

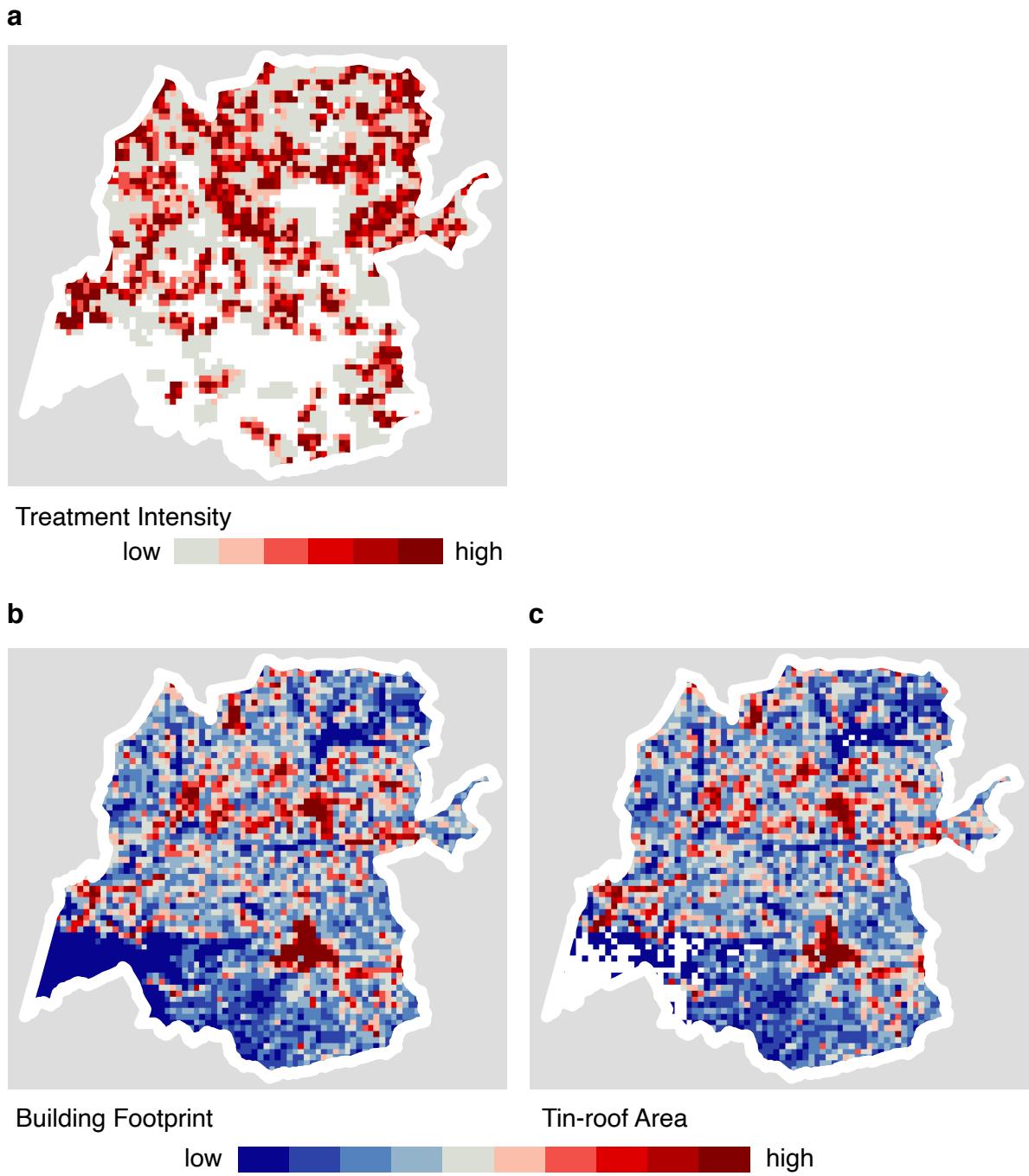


Figure 2: Maps of the treatment intensity (Panel a), building footprint (Panel b), and tin-roof area (Panel c). The white lines outline the sample area for the GiveDirectly trial.

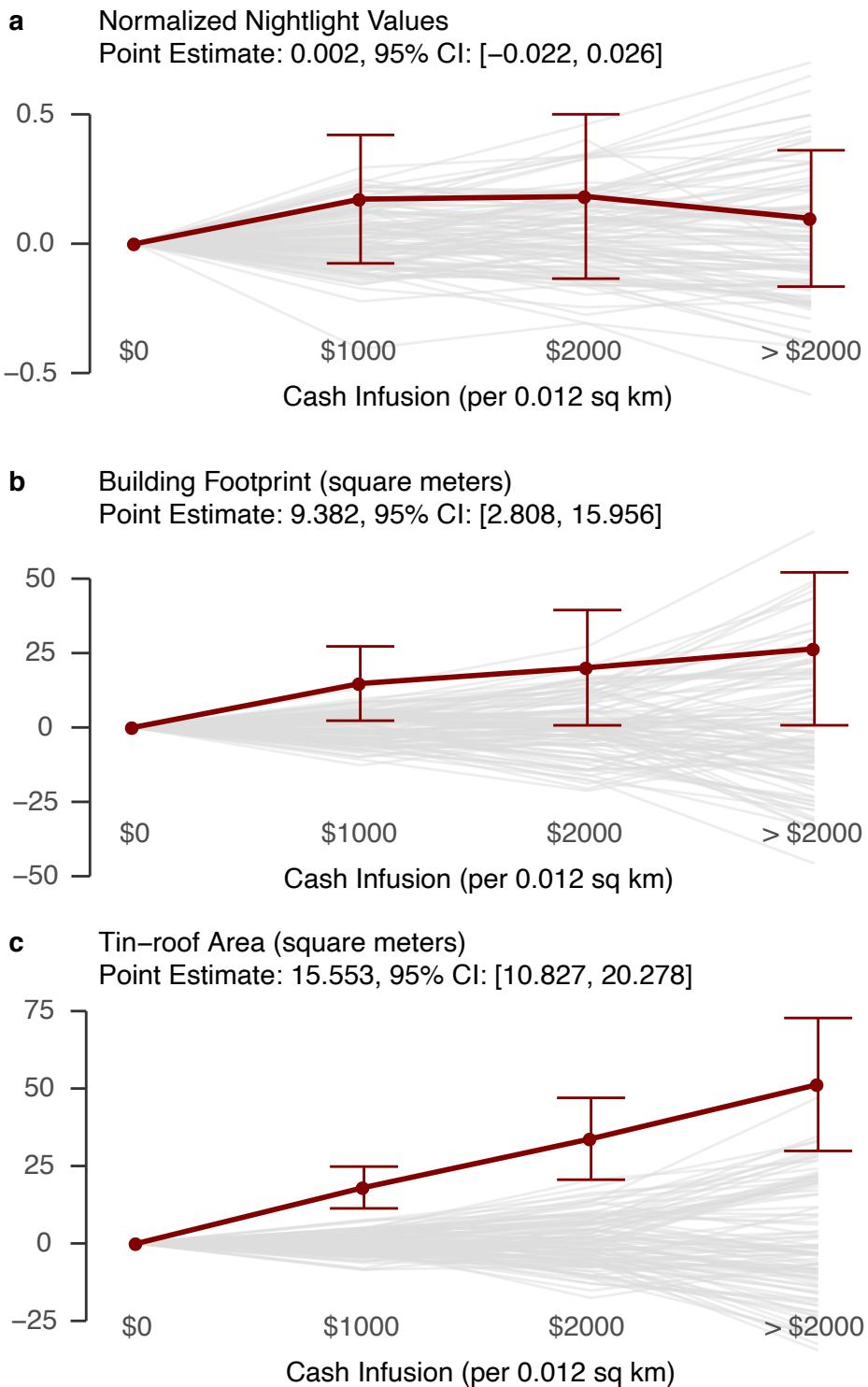
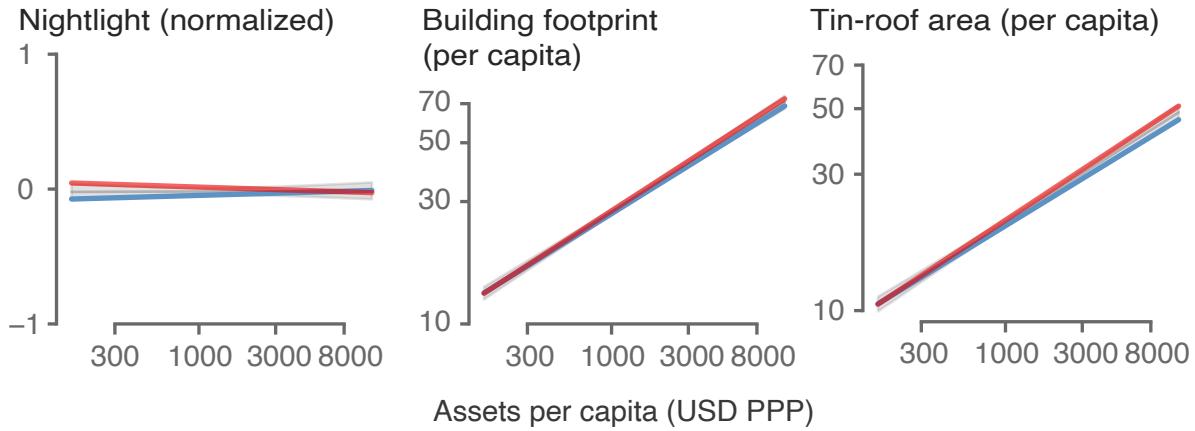


Figure 3: Average treatment effects of the GiveDirectly unconditional cash transfer on satellite derived variables. Gray lines indicate estimated treatment effects from 100 placebo simulations.

**a**

### Engel Curves



**b**

### Estimated versus Observed Treatment Effects on Assets Per Capita

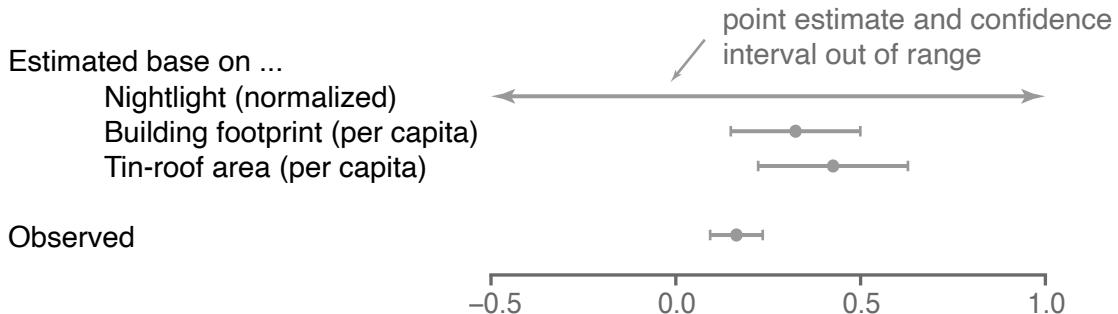


Figure 4: The Engel curves of nightlight, building footprint, and tin-roof area (Panel a), and the estimated versus observed treatment effects on assets per capita. The grey curves and grey regions in the left panels indicate the Engel curves (linear fit) and its 95% confidence intervals for the full eligible sample. Households who have received the treatment (red curve) do not have a different Engel curve from households who have not received the treatment (blue curve).

## References

1. L. Brune, D. Karlan, S. Kurdi, C. R. Udry, "Social Protection Amidst Social Upheaval: Examining the Impact of a Multi-Faceted Program for Ultra-Poor Households in Yemen", tech. rep. (National Bureau of Economic Research, 2020).
2. J. Puri, F. Rathinam, Challenges in real-world impact evaluations: Some learning on costs and timeliness. (2019).
3. A. Storeygard, Farther on down the road: transport costs, trade and urban growth in sub-Saharan Africa. *The Review of economic studies* **83**, 1263–1295 (2016).
4. A.-S. Isaksson, A. Kotsadam, Chinese aid and local corruption. *Journal of Public Economics* **159**, 146–159 (2018).
5. R. Jedwab, E. Kerby, A. Moradi, History, path dependence and development: Evidence from colonial railways, settlers and cities in Kenya. *The Economic Journal* **127**, 1467–1494 (2017).
6. J. Fenske, N. Kala, Climate and the slave trade. *Journal of Development Economics* **112**, 19–32 (2015).
7. T. Besley, M. Reynal-Querol, The legacy of historical conflict: Evidence from Africa. *American Political Science Review* **108**, 319–336 (2014).
8. Y. S. Lee, International isolation and regional inequality: Evidence from sanctions on North Korea. *Journal of Urban Economics* **103**, 34–51 (2018).
9. H. Bleakley, J. Lin, Portage and path dependence. *The quarterly journal of economics* **127**, 587–644 (2012).
10. S. Michalopoulos, E. Papaioannou, Pre-colonial ethnic institutions and contemporary African development. *Econometrica* **81**, 113–152 (2013).
11. S. Michalopoulos, E. Papaioannou, National institutions and subnational development in Africa. *The Quarterly journal of economics* **129**, 151–213 (2014).
12. A. Alesina, S. Michalopoulos, E. Papaioannou, Ethnic inequality. *Journal of Political Economy* **124**, 428–488 (2016).
13. R. Hodler, P. A. Raschky, Regional favoritism. *The Quarterly Journal of Economics* **129**, 995–1033 (2014).

14. S. Bazzi, A. Gaduh, A. D. Rothenberg, M. Wong, Skill transferability, migration, and development: Evidence from population resettlement in Indonesia. *American Economic Review* **106**, 2658–98 (2016).
15. N. Gennaioli, R. La Porta, F. L. De Silanes, A. Shleifer, Growth in regions. *Journal of Economic Growth* **19**, 259–309 (2014).
16. M. Lipscomb, A. M. Mobarak, Decentralization and pollution spillovers: evidence from the re-drawing of county borders in Brazil. *The Review of Economic Studies* **84**, 464–502 (2016).
17. J. V. Henderson, T. Squires, A. Storeygard, D. Weil, The global distribution of economic activity: Nature, history, and the role of trade. *The Quarterly Journal of Economics* **133**, 357–406 (2017).
18. M. Gonzalez-Navarro, M. A. Turner, Subways and urban growth: Evidence from earth. *Journal of Urban Economics* **108**, 85–106 (2018).
19. A. Kocornik-Mina, T. K. J. McDermott, G. Michaels, F. Rauch, Flooded Cities. *American Economic Journal: Applied Economics* **12**, 35–66, (<https://www.aeaweb.org/articles?id=10.1257/app.20170066>) (Apr. 2020).
20. G. De Luca, R. Hodler, P. A. Raschky, M. Valsecchi, Ethnic favoritism: An axiom of politics? *Journal of Development Economics* **132**, 115–129 (2018).
21. M. L. Pinkovskiy, Growth discontinuities at borders. *Journal of Economic Growth* **22**, 145–192 (2017).
22. F. Campante, D. Yanagizawa-Drott, Long-range growth: economic development in the global network of air links. *The Quarterly Journal of Economics* **133**, 1395–1458 (2017).
23. N. Jean *et al.*, Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
24. J. V. Henderson, A. Storeygard, D. N. Weil, Measuring economic growth from outer space. *American economic review* **102**, 994–1028 (2012).
25. S. Michalopoulos, E. Papaioannou, Spatial patterns of development: A meso approach. *Annual Review of Economics* **10**, 383–410 (2018).
26. A. Young, The African growth miracle. *Journal of Political Economy* **120**, 696–739 (2012).
27. A. Tarozzi, A. Deaton, Using census and survey data to estimate poverty and inequality for small areas. *The review of economics and statistics* **91**, 773–792 (2009).

28. D. Atkin, B. Faber, T. Fally, M. Gonzalez-Navarro,  
 “A New Engel on Price Index and Welfare Estimation”, tech. rep.  
 (National Bureau of Economic Research, 2020).
29. B. Marx, T. M. Stoker, T. Suri, There is no free house: Ethnic patronage in a Kenyan slum.  
*American Economic Journal: Applied Economics* **11**, 36–70 (2019).
30. G. Michaels *et al.*,  
 “Planning Ahead for Better Neighborhoods: Long Run Evidence from Tanzania”,  
 Unpublished, 2019.
31. T. A. Kohler *et al.*, Greater post-Neolithic wealth disparities in Eurasia than in North America and Mesoamerica. *Nature* **551**, 619–622 (2017).
32. T. A. Kohler, R. Higgins, Quantifying household inequality in Early Pueblo villages.  
*Current Anthropology* **57**, 690–697 (2016).
33. T. A. Kohler, M. E. Smith,  
*Ten thousand years of inequality: the archaeology of wealth differences*  
 (University of Arizona Press, 2018).
34. C. Elbers, J. O. Lanjouw, P. Lanjouw, Micro-level estimation of poverty and inequality.  
*Econometrica* **71**, 355–364 (2003).
35. C. Yeh *et al.*, Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications* **11**, 1–11 (2020).
36. J. Blumenstock, G. Cadamuro, R. On,  
 Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).
37. J. E. Blumenstock, Fighting poverty with data. *Science* **353**, 753–754 (2016).
38. A. Head, M. Manguin, N. Tran, J. E. Blumenstock, presented at the ICTD, pp. 8–1.
39. R. Engstrom, J. Hersh, D. Newhouse, “Poverty from space: Using high-resolution satellite imagery for estimating economic well-being”, Working Paper (The World Bank, 2017).
40. B. Babenko, J. Hersh, D. Newhouse, A. Ramakrishnan, T. Swartz,  
 Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in mexico. *arXiv preprint arXiv:1711.06323* (2017).
41. D. Egger, J. Haushofer, E. Miguel, P. Niehaus, M. W. Walker,  
 “General equilibrium effects of cash transfers: experimental evidence from Kenya”, tech. rep.  
 (National Bureau of Economic Research, 2019).
42. *Google Static Maps*, <https://developers.google.com/maps/documentation/maps-static/intro>, accessed 6 May 2020.

43. K. He, G. Gkioxari, P. Dollár, R. Girshick,  
presented at the Proceedings of the IEEE international conference on computer vision,  
pp. 2961–2969.
44. COCO - *Common Objects in Contexts*, <http://cocodataset.org>, accessed 6 May 2020.
45. 2018 Open AI Tanzania Building Footprint Segmentation Challenge,  
<https://competitions.codalab.org/competitions/20100>, accessed 6 May 2020.
46. T. G. Conley, GMM estimation with cross sectional dependence.  
*Journal of econometrics* **92**, 1–45 (1999).
47. T. Conley,  
*Spatial econometrics. New Palgrave Dictionary of Economics*, eds Durlauf SN, Blume LE,  
2008.
48. S. M. Hsiang, Temperatures and cyclones strongly associated with economic production in  
the Caribbean and Central America.  
*Proceedings of the National Academy of sciences* **107**, 15367–15372 (2010).
49. F. Burlig, M. Woerman, *ARE 212 Section 10: Non-Standard Standard Errors II*,  
[https://static1.squarespace.com/static/558eff8ce4b023b6b855320a/t/573bd63745bf21da74c080a8/1463539276997/ARE\\_212\\_Section\\_10.pdf](https://static1.squarespace.com/static/558eff8ce4b023b6b855320a/t/573bd63745bf21da74c080a8/1463539276997/ARE_212_Section_10.pdf),  
accessed 10 May 2020.
50. K. Lee, E. Miguel, C. Wolfram,  
Experimental evidence on the economics of rural electrification.  
*Journal of Political Economy* **128**, 1523–1565 (2020).

# Appendices

## A Appendix Figures and Tables



Figure A.1: 10 randomly sampled prediction results of the deep learning model.

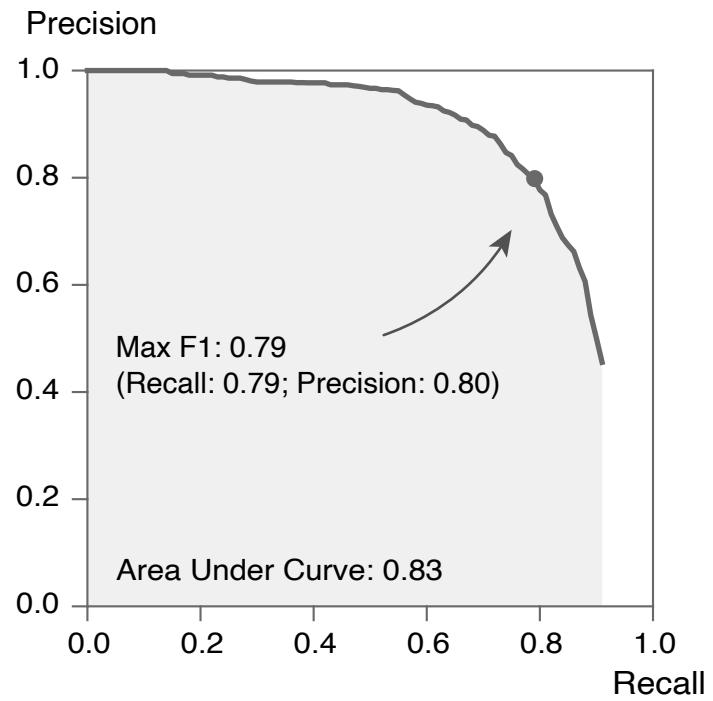


Figure A.2: The precision-recall curve of the Mask R-CNN model, evaluated on a random subset of 120 in-sample satellite images with 3-fold cross validation.

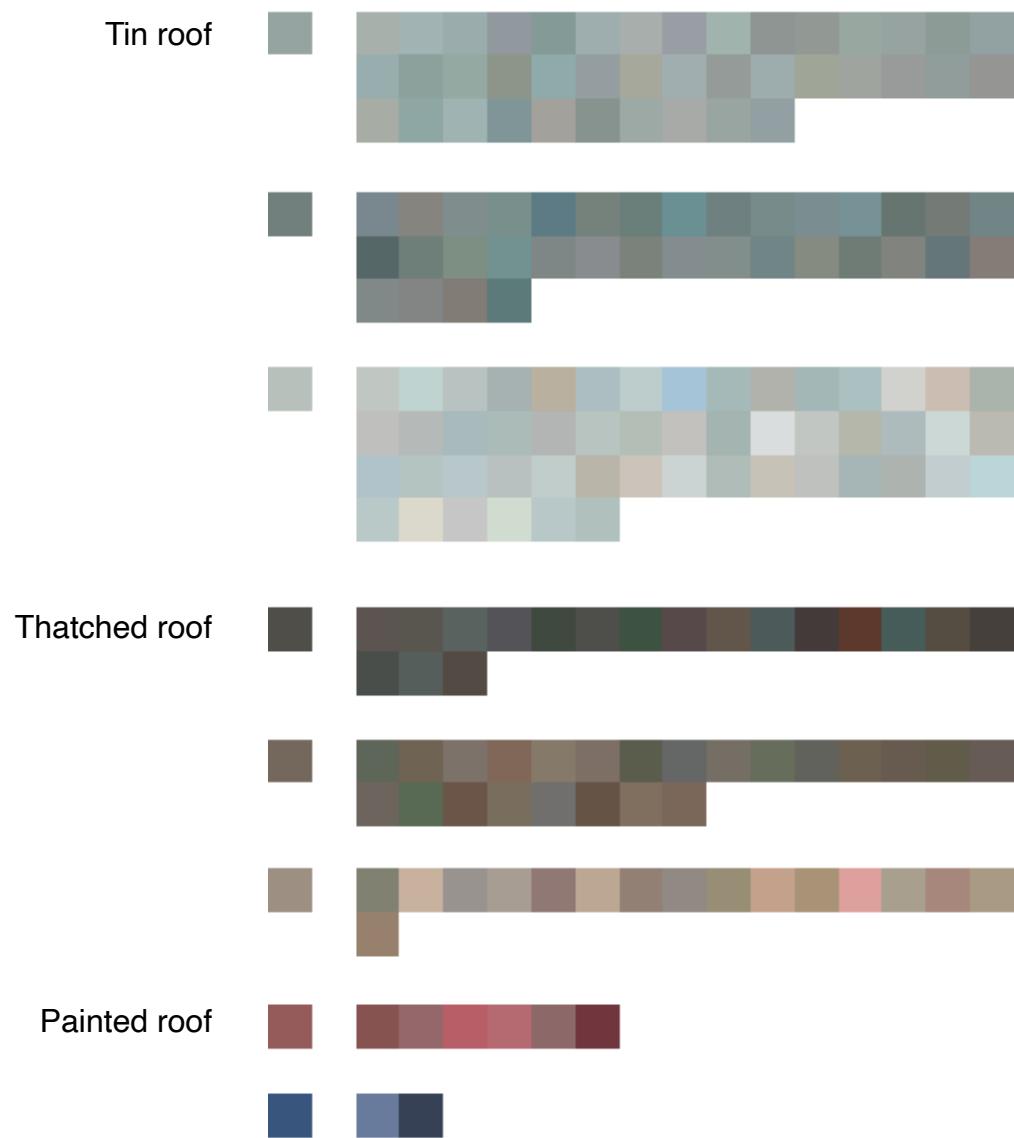
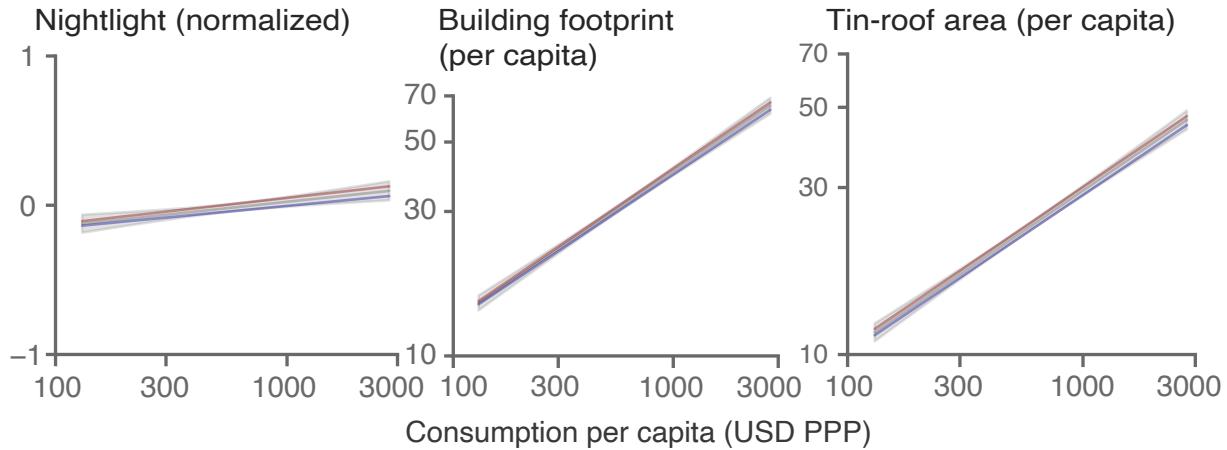


Figure A.3: The results of K-means clustering ( $K = 8$ ) of roof colors in the GiveDirectly sample area.

**a**

### Engel Curves



**b**

### Estimated versus Observed Treatment Effects on Consumption Per Capita

Estimated base on ...

- Nightlight (normalized)
- Building footprint (per capita)
- Tin-roof area (per capita)

Observed

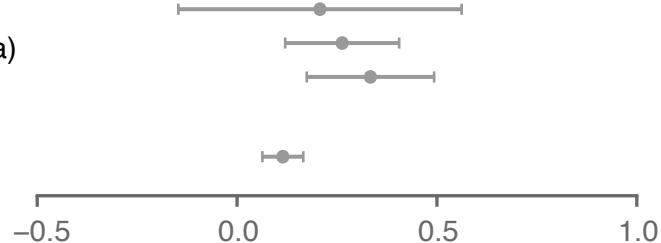


Figure A.4: The Engel curves of nightlight, building footprint, and tin-roof area (Panel a), and the estimated versus observed treatment effects on consumption per capita (Panel b). The grey curves and grey regions in the left panels indicate the Engel curves (linear fit) and its 95% confidence intervals for the full eligible sample. Households who have received the treatment (red curve) do not have a different Engel curve from households who have not received the treatment (blue curve).

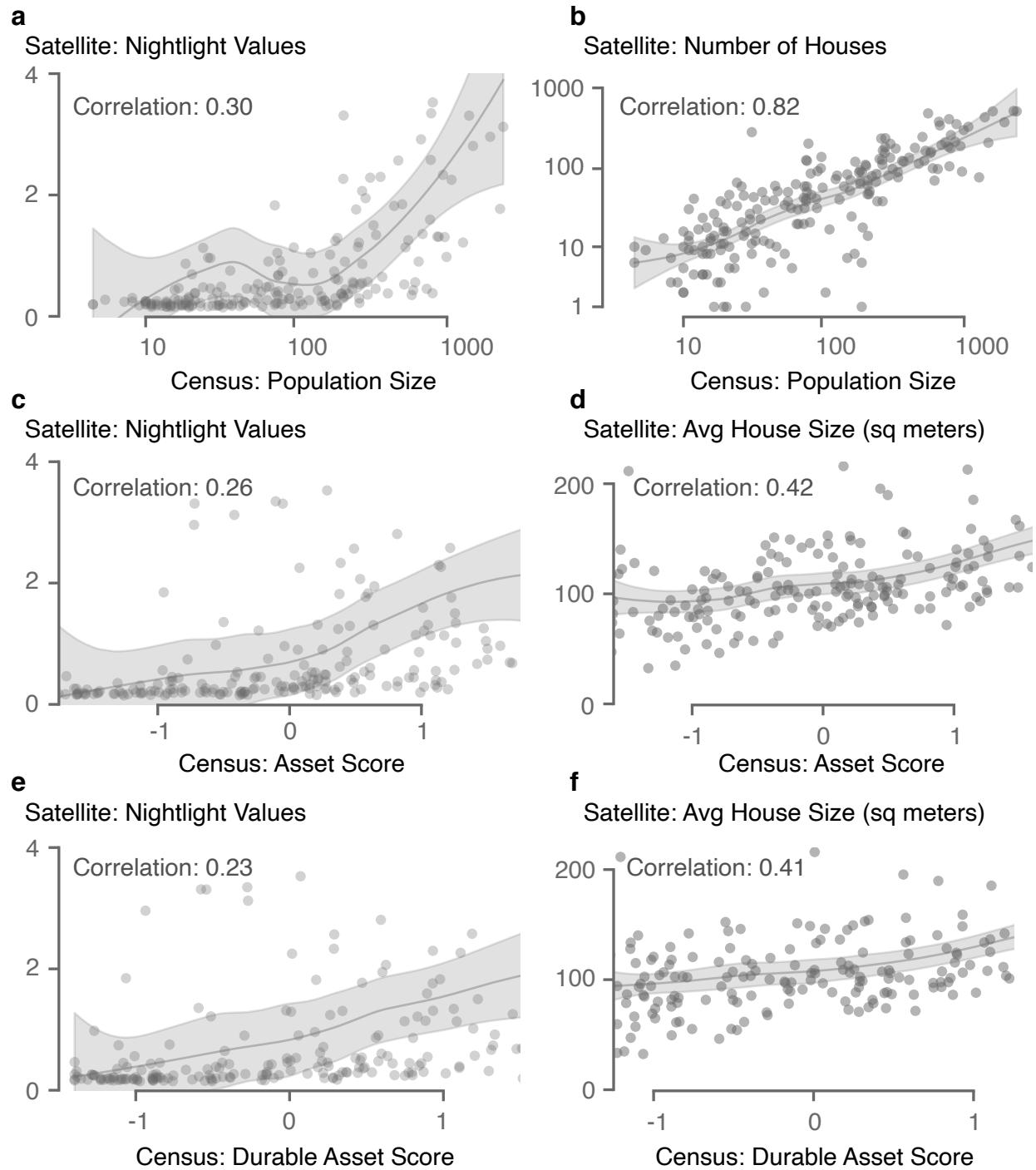


Figure A.5: Correlation between satellite-derived variables and census-derived variables in rural Mexico. Each point corresponds to a sampled rural locality in Mexico. Gray lines are estimated LOESS curves, along with 95% confidence intervals.

## B Data and Methods

The goal of this paper is to develop methods that generate inexpensive, remotely sensed proxies of living standards in developing countries, and demonstrate the feasibility of conducting causal inference with these proxy measures. Here, I describe my input data, including census, survey, and remote sensing data, and document the process of training a deep learning model (Mask R-CNN) to extract building footprint, and constructing measures of housing quality.

### B.1 Data

#### B.1.1 Census and Survey Data

**Census and Survey Data in the GiveDirectly Trial** In order to validate the remotely sensed living standard proxies, and demonstrate the feasibility of conducting causal inference with these measures, I acquire two data sets from the authors of the GiveDirectly randomized controlled trial in rural Kenya (1).

First, I acquire the baseline household census data in the GiveDirectly study area. Before the intervention, the authors first conducted a baseline household census in all the villages, which served as a sampling frame for the subsequent surveys and collected information on household eligibility status. The census identified 65,385 households with a total baseline population of 280,000 people in study villages. Consistent with the original paper, I use the sampling frame established by the baseline census in all the econometric analyses. From the census, I map the intensity of the treatment by combining information on the geo-location of the households, the households' eligibility status and the treatment status of the villages. Households that met both of the following requirements can receive the transfer: (1) they were eligible for the GiveDirectly unconditional cash transfers (i.e., they lived in a thatched-roof house at the time of the baseline census); and (2) they lived in a village that were randomized into the treatment group. The census was separately conducted by independent (non-GiveDirectly) enumerators to achieve consistency between the treatment and control villages, and the procedure was designed to mimic GiveDirectly's own censusing procedure. To minimize the impacts of the measurement errors generated by the GPS collection device, missing GPS coordinates or outliers were dropped and filled with village longitude and latitude means. Out of 65,385 households, 4 households had missing GPS coordinates, and 58 household coordinates were considered outliers (more than 2 kilometers away from the village averages). Additionally, I obtain the boundary shapefile of the study area (shown in Figure 2) from the authors.

Second, I acquire the endline survey data collected by the authors between May 2016 and June 2017. This sample is a subset of the census sample. The authors randomly surveyed 8 eligible households and 4 ineligible households per village, with a village containing about 100 households on average, of which around one third are eligible. In order to match with satellite imagery, I

drop households without GPS coordinates. My final analysis sample contains 7,987 households in 653 villages, of which 4,578 households are eligible for the transfer. From the endline survey data, I obtain high-quality measures of economic well-being, including total consumption and assets, constructed from a comprehensive consumption and expenditure survey. Household consumption expenditure is the annualized sum of total food consumption in the last 7 days, frequent purchases in the last month, and infrequent purchases over the last 12 months. Asset measures include livestock, transportation (bicycles, motorcycles, and cars), electronics, farm tools, furniture, other home goods, and lending/borrowing from formal/informal sources. Asset measures also include housing value, measured as the respondent's self-reported cost to build a home like theirs, and land value, measured as landholdings multiplied by the self-reported per-acre cost of land of similar quality in their village. I compute consumption or assets per capita by dividing the household consumption or asset values by the number of household members (including children). To minimize the influence of outliers, I winsorize consumption and asset per capita values at the 2.5% and 97.5% cut-off of the full (eligible and non-eligible) sample<sup>7</sup>.

**Census Data in Mexico** To further validate remotely sensed measures of housing quality, and assess the relationship between these measures and overall economic well-being, I conduct a validation exercise against the 2010 Population and Housing Census in Mexico (2).

I utilize the locality-level data set (Principales Resultados por Localidad, or ITER) in the 2010 census. I do not use household-level micro data in this paper, because they are not precisely geocoded. Each observation represents a locality, which is equivalent to a village in rural areas. I drop all urban localities (defined as having more than 2,500 residents) and focus only on rural localities. Additionally, I drop localities where the relevant asset measures are masked in the census to protect privacy (in very small localities), or where these measures are missing. Each rural locality is geocoded as a point. Most of the rural localities are small, isolated and surrounded by vegetation or open space, alleviating concerns that the 2010 census did not include the boundary of the rural localities. I match each locality to satellite images that cover an area of roughly  $1 \times 1$  km, with the locality coordinate at the center. To avoid covering neighboring urban or rural localities in the satellite images, I exclude rural localities that are closer than 1.1 km (0.01 degree) from other rural localities, or 11.1 km (0.1 degree) from urban localities. Finally, to reduce computation, I randomly sampled 200 rural localities for the validation exercise. Google Static Maps does not have satellite image coverage for 3 of the 200 sampled localities and they are thus dropped from the sample.

In line with prior literature (3, 4), I use an asset index to measure wealth. The asset index is computed as the first principal component of multiple variables, each measuring the percentage of households in the locality who own a given type of asset. Assets are grouped into three categories: durable good (radio, TV, refrigerator, washer, car, computer, telephone, cell phone, and internet),

---

<sup>7</sup>In the data set that I acquired, the household consumption values have already been winsorized at the 99.64% cut-off to reduce the effects of outliers.

housing (cement floor, house with  $\geq 2$  bedrooms, house with  $\geq 3$  rooms), and public good (toilet, electricity, piped water, and drainage). An asset index is calculated for each category, and an overall asset index is calculated by pooling all the categories. When necessary, I flip the signs of the asset indices, such that a higher score indicates more wealth.

### B.1.2 Daytime High-resolution Satellite Images

The input high-resolution satellite images are from Google Static Maps (5). These images have been pre-processed (e.g., to remove clouds) and geo-referenced and they come from a variety of sources such as Maxar (formerly DigitalGlobe) and Airbus. Unlike other remote sensing data sets, these images only contain the RGB (red, green, blue) bands, and no information in other wavelengths.

The input satellite images are at a spatial resolution of about 30 cm per pixel (on equator). More precisely, the images are at zoom level 19 in the XYZ tiles system, a convention that Google Static Maps uses to store and display satellite images at different scales. Users can retrieve an image of  $640 \times 640$  pixels via querying the API with the geo-coordinate of the center of the image. Each Google Static Maps image cover an area of around  $190 \times 190$  meters at zoom level 19. The XYZ tiles system is based on the Web Mercator projection system, which induces large area distortion in high-latitude areas. Whenever necessary, I correct for area distortion during the post processing steps.

Google Static Maps uses satellite images from different commercial providers and different time periods, and seamlessly mosaic images together. Most images are taken fairly recently. However, no exact timestamps are available for the images.

To evaluate the GiveDirectly randomized controlled trial in Kenya, I download a set of satellite images covering all of the Siaya county in Western Kenya, where the study took place. The images were retrieved from the Google Static Maps API between Feb 19 and Feb 21, 2020. Given that the GiveDirectly cash transfers were given out from mid 2014 to early 2017, it is safe to assume that most of these satellite images are taken after the intervention. This is consistent with the fact that I observe strong treatment effects.

To conduct the validation exercise against the 2010 census in Mexico, I download a collection of satellite images covering about an area of  $1 \times 1$  km, for each rural locality in the analysis sample. More precisely, I download a grid of  $5 \times 5$  satellite images (each with a size of  $640 \times 640$  pixels) with the geo-coordinate of the locality being the center of the image grid. The images were retrieved from the Google Static Maps API on October 10, 2019. The time when the satellite images are taken is quite different from the time of the census, likely reducing the correlation between satellite- and census-based measures.

### B.1.3 In-sample Building Footprint Annotations

I create in-sample building footprint annotations to objectively and quantitatively evaluate the predictive performance of the deep learning model. Among the 71,012 satellite images that cover all of the Siaya county in Kenya, I randomly sample 120 images for annotation. Among the 197 rural localities in Mexico, I randomly sample 8 localities and annotated all the 199 satellite images in these localities.

I use the supervisely image annotation web platform to create annotations. On any given image, I outline the boundaries of all the instances of buildings on the image. Buildings that border each other are annotated as separate instances, if there are reasons to believe that they are separate structures (e.g., if they appear to use different roof materials). Half-finished buildings are included in the annotations, although they are fairly rare in the analysis sample.

Some measurement errors are expected to arise from the annotation exercise, but their magnitude is small. First, the Google Static Maps logo block 1.05% of the total area of any given image, and structures covered by the logos are not annotated. Second, only the visible parts of the buildings are annotated, but a very small part of some buildings may be partially occluded by trees. Third, as all the satellite images are processed as “chips” with a size of  $640 \times 640$  pixels, a small proportion of structures may be on the edges of the chips and thus annotated as “partial” structures on multiple chips. Finally, the annotation accuracy (and potentially prediction accuracy) may be different across buildings with different roof materials. In particular, some study participants in the GiveDirectly trial live in thatched-roof houses. These structures tend to be harder to identify for human annotators than metal-roof houses, because they are typically smaller, not as reflective, and may sometimes be mistaken for a tree in overhead imagery.

Almost all the buildings in my analysis sample are single-storey structures that are fairly isolated. For this reason, building footprint may be a good approximation of overall square footage of the buildings, and the number of structures tend to be well correlated with population density. This may apply to other rural settings. However, the existence of apartment complexes, and a large number of multi-storey buildings will present challenges to generalizing the methods in this paper for more urban contexts.

### B.1.4 Nightlight

Satellites from the United States Air Force Defense Meteorological Satellite Program (DMSP) have been recording the intensity of Earth-based lights with their Operational Linescan System (OLS) sensors, creating an important remote sensing data set that are widely used to proxy economic development (6). Beginning in 2014, the National Oceanic and Atmospheric Administration’s (NOAA) National Geophysical Data Center (NGDC) started to provide a separate nighttime lights data set collected from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB). The VIIRS-DNB data are produced at a spatial resolution of 15 arc seconds (approximately 463

meters at the equator), twice the resolution of the preceding DMSP-OLS product. It is considered superior to DMSP-OLS because it preserves more spatial details, has a lower detection limit and displays no saturation on bright lights (which is common in dense urban areas) (7). Most of the empirical economic studies (6) and related work (3) have been based on the DMSP-OLS product because of its more extensive temporal coverage. Throughout this paper, I use the VIIRS-DNB product. I choose the year 2019 to minimize the temporal differences between daytime Google Static Maps images and nightlight data products, and to conduct fairer comparisons. Specifically, I download the stray light corrected version of the monthly average radiance composite images hosted on Google Earth Engine (8). I aggregate over all the monthly observations in 2019 by taking the mean.

The VIIRS-DNB data product excludes areas impacted by cloud cover and correct for stray light (9). However, it has not been filtered to screen out lights from aurora, fires, boats, and other temporal lights. Additionally, lights are not separated from background (non-light) values. These methods are currently under development (8).

For the causal inference or validation exercises, I link the nightlights raster data set to either household coordinates (collected in the GiveDirectly census or survey in Kenya) or locality coordinates (collected in the 2010 census in Mexico). Specifically, I assign the nightlight values in a given pixel to all the households or localities in that pixel. I winsorize nightlight values at 99 percentile to reduce the influence of outliers on the overall result, and in some cases, normalize the distribution to improve interpretability.

## B.2 Methods

Deep learning models have proven to be immensely successful in establishing new benchmarks on a wide set of computer vision tasks, such as image recognition and segmentation, that have traditionally been challenging to automate (10). The method presented in this paper builds on established computer vision algorithms to analyze satellite images and study the relationship between remotely sensed proxies and measures of economic development collected in field surveys. Specifically, I focus on one of the most salient features in satellite images that carry strong signals about wealth: housing. I use a state-of-the-art instance segmentation model, Mask R-CNN (11), to perform a building footprint instance segmentation task. Here, I provide details on the training schedule, the hyper-parameter settings of the model, and the post-processing steps that generate measures for housing quality.

### B.2.1 Mask R-CNN

**Intuition** Instance segmentation is a computer vision task where the algorithm identifies the boundaries of all the instances belonging to a category of interest (in this paper, buildings) on input images. Mask R-CNN is a model that first generates potential proposals of bounding boxes

that contain the objects of interest; and then predicts what class the objects belong to and outlines the boundaries of the identified objects. Consider the following set-up

$$Y = f_\theta(X) \tag{7}$$

In this formulation,  $X$  denotes the input matrix to the model,  $Y$  denotes a 4-tuple of the output matrices, and  $f_\theta$  denotes the deep learning model with the parameters  $\theta$ . For a given input image,  $X$  is a  $H \times W \times C$  ( $H$ : height;  $W$ : width;  $C$ : color) matrix that represents the raw pixel values of the RGB (Red, Green, Blue) bands, with each element taking on an integer value between 0 to 255. For every image, the model predicts a maximum of 100 instances. Assuming that there are  $N$  instances detected in the image, then  $Y$  is a 4-tuple of  $(B, L, S, M)$ , where  $B$  is a matrix of size  $N \times 4$  representing  $N$  bounding boxes;  $L$  is a vector of length  $N$  representing the predicted labels for each instance (with 1 denoting buildings);  $S$  is a vector of length  $N$  representing the scores of each prediction (ranging from 0 to 1); and  $M$  is a matrix of size  $H \times W \times N$  representing predicted pixel masks for all the instances<sup>8</sup>.

During the training phase, the model is shown a large number of  $(X, Y)$  pairs in the training data. The model parameters  $\theta$  are optimized to predict instance segmentation results  $Y$  based on input images  $X$ , such that the difference between the predictions and the “ground truth” is minimized.

**Implementation Details** The model is based on the official PyTorch implementation of Mask R-CNN. I use ResNet50 with the Feature Pyramid Networks as the backbone of the model. The model is trained with a learning rate of  $5 \times 10^{-4}$  and a batch size of 10. Optimization is conducted with the Adam optimizer.

The official PyTorch implementation of the original Mask R-CNN model cannot handle training images that are empty (i.e., images that do not contain any buildings). This may be particular problematic in the context of building detection and segmentation in developing countries, because the population density is low and a majority of the images will be empty. Discarding these empty images in training time may cause undesirable shifts in the data distribution that lower model performance. To address this issue, I create a “placeholder” class (which is later dropped) that corresponds to a small gray patch, and paste that onto all the empty images so that they can be included in the training data.

### B.2.2 Training Schedule

One of the main challenges in training building footprint instance segmentation models in developing countries is that training data are sparse. Most mainstream data sets on building footprint, such

---

<sup>8</sup>Elements in  $M$  take on continuous values between 0 and 1. Following convention, I perverse the continuous values in training time, but threshold  $M$  at 0.5 to obtain Boolean masks when conducting inference.

as SpaceNet and Inria Aerial, consist of images in metropolitan cities in developed countries (e.g., Chicago, Las Vegas, Paris and Shanghai). These urban landscapes are dramatically different from most rural scenes in my analysis sample. Additionally, both SpaceNet and Inria Aerial provide semantic segmentation labels, and do not annotate buildings that border each other separately.

To overcome these challenges, I train the Mask R-CNN model with a multi-step process, utilizing both publicly available data sets and data sets created specifically for this paper. First, I pre-train the model on Open AI Tanzania, a collection of overhead images taken by consumer drones in Zanzibar, Tanzania, that are annotated with building footprints (12). These images are more representative of the rural or peri-urban scenes in a developing country context, in terms of the distribution of the density, sizes and heights of the buildings. Second, I create in-sample annotations by randomly sampling from the set of images that predictions will be made on, and annotating them (see Section B.1.3 for a more detailed description). This ensures that training images and inference images belong to the same data distribution. Because the set of in-sample annotations is small, I rely on transfer learning (i.e., pre-training on larger data sets) to prevent overfitting.

Throughout the training process, I conduct extensive data augmentation to increase the transferability of the model from one data set to another. I randomly flip the training images horizontally and vertically, randomly jitter the brightness, contrast, saturation, and hue of the images, and for Open AI Tanzania, randomly blur and crop the images.

Below, I describe in more detail the training schedule and the data used.

**1. COCO (Common Objects in Context)** The model is first pre-trained with the COCO (Common Objects in Context) data set, a large-scale natural image data set containing 80 object categories and around 1.5 million object instances (13). Despite the fact that input images and object categories in COCO are different from target satellite images, pre-training the model with a large-scale dataset often provides meaningful performance gains, even when the model is later transferred across domains. I load the COCO pre-trained model provided in PyTorch. Because COCO has a different set of object categories, I re-initialized weights in the last neural network layers used for predicting bounding boxes and masks.

**2. Open AI Tanzania** After the first step, the model is fine-tuned on the Open AI Tanzania building footprint segmentation data set, a collection of high-resolution aerial imagery collected by consumer drones in Zanzibar, Tanzania (12).

All the buildings in the drone images are identified, outlined and classified into three categories (completed building, unfinished building, and foundation) by human annotators. This somewhat unusual categorization is due to the fact that there exist a large number of unfinished structures in Zanzibar. Most input satellite images in this paper contain very few unfinished structures, so I do not preserve it as a separate category. I collapse the first two categories into buildings and drop the third category.

The native resolution of the drone images is 7cm. I down-resolution the images to about 30cm to match with the resolution of the target satellite images. The images are provided as large tiles, which cannot be fed to the deep learning model directly, so I randomly sample smaller chips from the large tiles (with an oversampling ratio of 3, meaning that any area in the sample is sampled 3 times in expectation). This serves as a form of data augmentation and takes advantage of the fact that the raw input images and annotations cover large contiguous areas.

In training time, 90% of the sampled 4,034 chips are used for training, and the remaining 10% for validation. In order to guard against overfitting, and choose the best model, in each epoch, I evaluate the performance of the model with the validate set, using average precision with an Intersection over Union (IoU) cutoff of 0.5 as the main evaluation metric. The model is trained for 50 epochs, and the best model (at epoch 43) is saved and loaded in subsequent steps.

### **3. In-sample Annotations in Mexico; Supplementary Images in Tanzania and Kenya**

Next, the model is fine-tuned on a set of 587 annotated high-resolution satellite images, including

- 199 Google Static Maps images in Mexico that are randomly drawn from my analysis sample (see Section B.1.1 and B.1.2 for details on the construction of the sample) and annotated with building footprints (see Section B.1.3 for details on annotation procedures).
- 200 Google Earth Pro images in Tanzania that are randomly sampled from several peri-urban neighborhoods in Dar es Salaam. Unlike most Google Static Maps images, these images are historical images from 2003 to 2018, and generally have a lower image quality. They are sometimes covered by cloud, and buildings masked by clouds are not annotated. These annotations were originally created for another project, but implemented in a consistent manner as our in-sample annotations.
- 188 Google Static Maps images in Kenya that are sampled from the GiveDirectly study area. These images were originally downloaded in 2018 and annotated in the same manner as described in Section B.1.3. Updates in both the main analysis and the satellite images in 2020 rendered these downloaded images obsolete, and they were superseded by newly created in-sample annotations. I continue to use them in this pre-training step.

These images are pooled and randomly split into a training set (90%) and a validation set (10%). The model is fine-tuned for 25 epochs, achieving the best performance at epoch 17. For the validation exercise in Mexico, the best model here is used for inference; for the evaluation of the GiveDirectly trial, I fine-tune the best model further in the next step.

### **4. In-sample Annotations in Siaya, Kenya**

For the evaluation of the GiveDirectly trial, the final step is to fine-tune the model on in-sample annotations in Siaya, Kenya. Section B.1.1 and B.1.2 describe how the sample is constructed and Section B.1.3 describes the procedure for

annotation. Here, the model is trained on 90% of the images for 25 epochs, and evaluated with the 10% held out set, which determines that the best performance is achieved at epoch 15. Inference on satellite images in Siaya, Kenya is based on the best model estimated in this step.

### B.2.3 Evaluation Metrics

I provide quantitative performance evaluation of the Mask R-CNN model. The performance metrics reported in this section is representative of that on the entire GiveDirectly analysis sample, because the evaluation is conducted on an annotated random subset of the GiveDirectly images (see Section B.1.3 for details).

Figure A.2 shows the precision-recall curve on the in-sample annotations in Siaya, Kenya, computed with three-fold cross validation. First, I load the best model in Section B.2.2, Step 3. Then, I train on 67% of the sample for 15 epochs, and obtain predictions for the remaining 33%, a process repeated to generate predictions for all the images in the sample. Finally, I calculate the precision-recall curve based on the entire set of predictions. The Mask R-CNN model generates a confidence score for every predicted instance, and by varying the confidence score cutoff, one could adjust their (potentially different) tolerance for false positive rates or false negative rates. The precision ( $1 - FalsePositiveRate$ ) recall ( $1 - FalseNegativeRate$ ) curve displays how the model performs under different score cutoffs. For example, if all the instances, regardless of their confidence scores, are “accepted” (on the rightmost point of the curve in Figure A.2), one would expect the false negative rates to be low, and the false positive rates to be high (most true instances are found, but many predicted instances are wrong); on the other hand, if only a small set of instances with extremely high confidence scores are “accepted” (on the leftmost point of the curve in Figure A.2), one would expect the false negative rates to be high, and the false positive rates to be low (most predicted instances are correct, but many true instances are missed). Here, a true positive instance is defined as the predicted pixel mask and the ground truth pixel mask having sufficient overlap (more precisely, the intersection of the two masks is more than 50% of the union of the two masks).

The area under the precision-recall curve (i.e., average precision) is used for model selection as described in Section B.2.2. The precision-recall curve shown in Figure A.2 is not complete—the rightmost part of the curve is not shown because the Mask R-CNN model default to not outputting predictions with a confidence score that is lower than 0.05, for efficiency reasons. I do not expect this to systematically impact model selection.

To determine the confidence score cutoff, below which predicted instances are dropped, I maximize F1, the harmonic mean of precision and recall. Using this criterion, I choose the confidence score cutoff of 0.73, resulting in an F1 score of 0.79, with precision being 0.80 and recall being 0.79 (Figure A.2). As a reference point, the No. 1 winner in the 2nd SpaceNet building footprint extraction competition reported an F1 score of 0.70 (averaged over Las Vegas, Shanghai, Paris and Khartoum). In particular, the F1 score in a developing country context (Khartoum) is notably

lower, at 0.54. The scores on different data sets are not directly comparable, and the performance difference may be due to the fact that building footprint segmentation in rural, less complex scenes are relatively easier than in modern cities. Future work may conduct comprehensive ablation studies to determine the optimal training schedule, or compare the Mask R-CNN model to others, but that is beyond the scope of the present study.

#### B.2.4 Post-processing Steps

For each predicted building instance, I convert the Boolean pixel mask to a polygon, and simplify the polygon with the Douglas-Peucker algorithm with a pixel tolerance of 3. From these polygons, I extract several observable properties that may be good indicators of housing quality.

**Roof Material** I extract an “average” roof color for every building, as shown in Figure 1 (Panel e). I overlay a predicted building polygon with the raw input satellite image, take all the pixels in the polygon, and average over the pixel values in the RGB (Red, Green, Blue) channels respectively. Roof color is often a good proxy for roof materials, and has been used by prior work to measure housing quality (14).

To improve the interpretability of roof colors, I cluster these roof color vectors into 8 groups by running the K-means clustering algorithm. Since the Euclidean distances between color vectors in the RGB color space does not properly reflect perceptual differences, I first project all the RGB color vectors to the CIELAB color space, and then perform K-means clustering in that space. Figure A.3 show the 8 color clusters formed, and I group them further into tin, thatched, and painted roof materials. The color on the left represents the “average” color, and the colors on the right represent a random sample of all the colors belonging to this cluster. The number of squares on the right is proportional to the number of observations belonging to the clusters.

**Building Footprint** I compute the area covered by each predicted building polygon and convert the unit to square meters. To reduce the effects of outliers, I winsorize this variable at the 1 and 99 percentile. If matched with population density or household size, I calculate building footprint per capita through dividing overall building footprint by the number of residents or the number of household members (both including children).

## References

1. D. Egger, J. Haushofer, E. Miguel, P. Niehaus, M. W. Walker,  
“General equilibrium effects of cash transfers: experimental evidence from Kenya”, tech. rep.  
(National Bureau of Economic Research, 2019).
2. *2010 Population and Housing Census of Mexico*,  
<https://www.inegi.org.mx/programas/ccpv/2010/default.html>, accessed 5 May 2020.
3. N. Jean *et al.*, Combining satellite imagery and machine learning to predict poverty.  
*Science* **353**, 790–794 (2016).
4. J. Blumenstock, G. Cadamuro, R. On,  
Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).
5. *Google Static Maps*, <https://developers.google.com/maps/documentation/maps-static/intro>,  
accessed 6 May 2020.
6. J. V. Henderson, A. Storeygard, D. N. Weil, Measuring economic growth from outer space.  
*American economic review* **102**, 994–1028 (2012).
7. C. D. Elvidge, K. E. Baugh, M. Zhizhin, F.-C. Hsu,  
Why VIIRS data are superior to DMSP for mapping nighttime lights.  
*Proceedings of the Asia-Pacific Advanced Network* **35** (2013).
8. *VIIRS Stray Light Corrected Nighttime Day/Night Band Composites Version 1*,  
[https://developers.google.com/earth-engine/datasets/catalog/NOAA\\_VIIRS\\_DNB\\_MONTHLY\\_V1\\_VCMSCFG](https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_MONTHLY_V1_VCMSCFG),  
accessed 6 May 2020.
9. S. Mills, S. Weiss, C. Liang, presented at the Earth Observing Systems XVIII, vol. 8866,  
88661P.
10. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *nature* **521**, 436–444 (2015).
11. K. He, G. Gkioxari, P. Dollár, R. Girshick,  
presented at the Proceedings of the IEEE international conference on computer vision,  
pp. 2961–2969.
12. *2018 Open AI Tanzania Building Footprint Segmentation Challenge*,  
<https://competitions.codalab.org/competitions/20100>, accessed 6 May 2020.
13. *COCO - Common Objects in Contexts*, <http://cocodataset.org>, accessed 6 May 2020.
14. B. Marx, T. M. Stoker, T. Suri, There is no free house: Ethnic patronage in a Kenyan slum.  
*American Economic Journal: Applied Economics* **11**, 36–70 (2019).