Evaluating Development Aid Programs with Satellite Imagery and Deep Learning

Luna Yue Huang*

This Version: February 9, 2021

Abstract

High-resolution daytime satellite imagery and deep learning methods promise to dramatically reduce the costs of program evaluation in international development. Here, I provide the first evidence that we could measure how programs change housing quality and living standards entirely remotely, incurring virtually no costs. As a proof of concept, I evaluated a development aid program in rural Kenya, and observed statistically significant and economically sizeable increases in building footprint and roof quality. With an Engel curve approach, I inferred overall program effects on recipients' living standards, and obtained consistent results with extensive in-person surveys.

*University of California, Berkeley; Doctoral Fellow at the Global Policy Lab. Contact: 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, Supreet Kaur, 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 GiveDirectly team, for generously sharing the dataset with me and responding to my various inquiries. All errors are my own.

Introduction

New technologies promise to disrupt the traditional paradigm in distributing, targeting and evaluating development aid. Charities that deliver direct cash transfers to families in need via mobile money are able to dramatically reduce operational costs on the ground [1]. Economists have made tremendous progress in leveraging the ever growing repository of high-resolution satellite imagery, and new deep learning techniques to precisely identify the poor and target aid [2–9]. These technologies can similarly allow economists to conduct timely and inexpensive data collection and evaluate program impacts without having to coordinate costly and logistically challenging field work.

Rigorous program evaluation based on in-person household surveys has formed the basis of the modern approach of fighting against global poverty [10, 11]. Household surveys—typically comprehensive questionnaires containing hundreds of questions that touch every aspect of people's financial lives—help generate crucial insights into the effectiveness of different development aid programs, and provide input for evidence-based policy making. However, they can sometimes be prohibitively expensive to conduct. Studies funded by major donors cost 0.5 million US dollars on average (according to the World Bank) and up to 1.5 million US dollars (according to USAID) [12]. Unanticipated events, such as political unrest and pandemics, often disrupt field surveys, leaving crucial data missing [13].

The most widely used remotely sensed measure of economic development is nighttime light intensity (hereafter "night light"), which captures the amount of light emitted from Earth at night, and is highly correlated with Gross Domestic Product (GDP), as well as subnational economic development [14–16]. However, the night light data show poor sensitivity in less developed and rural areas [2], presumably because of low electrification rates—for example, from 1992 to 2008, 99.73% of pixels were completely unlit in Madagascar, 99.47% in Mozambique, and this is representative of low-income countries [14]. This makes the data less useful for studying the very target of many development aid programs—people living under the international poverty line. Additionally, the low spatial granularity of night light prevents it from being used to evaluate programs that generate fine spatial variations, including most of the randomized controlled trials, the gold standard for program evaluation.

Here, I propose a framework to measure housing quality from high-resolution daytime satellite imagery with a state-of-the-art deep learning model, Mask R-CNN [17]. With a proof-of-concept experiment in rural Kenya [18], I show how one could directly observe the impacts of development aid programs on housing quality, and infer the impacts on living standards with a canonical micro-economic concept, the Engel curve [19].

Housing quality promises to be a better proxy for economic development than night light. The visual appearance of buildings in satellite imagery (for example, roof color, roof reflectance, building footprint, and the geometric alignment between neighboring buildings) contains rich information about the quality of housing [20–22]. These metrics remain sensitive and measurable, even in poor

and rural communities with low electrification rates. Expenditure on housing accounts for a sizable 10–20% of people's total expenditure [23], and housing quality is generally a strong correlate of socio-economic status.

As a proof of concept, I evaluated a randomized controlled trial conducted in 2014–2017 in 653 villages in rural Kenya [18]. GiveDirectly, a US charity, implemented a program to send unconditional cash transfers to rural households via mobile money, if they met the eligibility criteria of living under a thatched roof (a low quality roof material, often indicating poor living conditions). Each recipient household would receive $1,000—equivalent to about 75% of their annual household expenditure—in lump sum, and could spend it however they want to. To evaluate the effectiveness of the program, GiveDirectly randomly selected about half of the 653 villages as the treatment group, where eligible households (about 1/3 of the population) received transfers, and used the rest as the control group. (In effect, the trial employed a more sophisticated two-tier randomization design, which is not directly relevant to the statistical analysis in this study.) The authors conducted extensive household surveys after the distribution of the transfers, and comprehensively measured program impacts.

Results

Mapping Treatment Intensity and Housing Quality. To evaluate program impacts, I first constructed a map that shows the intensity of the policy or development aid program (hereafter "treatment") in different geographical units (for example, raster grid cells). This can be derived from spatially explicit program implementation records, which document where the program was administered. Because of the high resolution of housing quality metrics, one could study programs that induced extremely fine spatial variation—for example, household-randomized trials. Importantly, the variation in treatment intensity has to be either random (if induced by an experiment) or as good as random (in a natural experiment setting), as is the case for any credible applied econometric study.

For the GiveDirectly experiment, I constructed the treatment intensity map from a local census in 2014–2015, which surveyed all the 65,385 households living in the study area [18]. The census data record each household's geo-location, and whether they belong to the treatment (T), control (C), or out-of-sample (O) group (Figure 1a). Among the three groups, only the treatment households eventually received the cash transfer from GiveDirectly. The control households were randomized into not receiving the transfer, whereas the out-of-sample households were never eligible to participate in the program. I laid out a regular grid, and counted the number of treatment households in each grid cell (Figure 1b). As every transfer was roughly USD 1,000, this variable can be interpreted as the amount of cash infusion (in $1,000) into a given grid cell, and is my preferred measure of treatment intensity (Figure 1e, Treatment Intensity).

Then, I measure housing quality in daytime satellite images with deep learning techniques. The

input images are from Google Static Maps [24]. They are likely from 2019, have a spatial resolution of about 30cm per pixel, and contain only the RGB (red, green, blue) bands (Figure 1c).

To segment buildings, I trained a state-of-the-art deep learning model, Mask R-CNN [17], on large, publicly available datasets such as COCO (Common Objects in Context) [25] and Open AI Tanzania [26], as well as a small annotated dataset, which were randomly sampled from all the input images (see Supplementary Materials B for details on model training). The model predictions are highly accurate, both quantitatively (Supplementary Figure S1) and qualitatively (Supplementary Figure S2). After post-processing, each predicted instance of buildings is represented by a polygon and a "representative" roof color (Figure 1d). The Mask R-CNN model conducts instance segmentation (as opposed to semantic segmentation), meaning that it is able to identify every building instance separately, even if they are adjacent to each other. As such, I could measure housing quality for each household, and evaluate program impacts at an unprecedentedly high spatial granularity.

I extracted two quality metrics for each building: the size of building footprint, and the type of roof material. The roofs were classified into three types: tin roof, thatched roof, and painted roof, based on their color profiles (Supplementary Figure S3). Compared to tin roofs, thatched roofs are generally of lower quality—they are cheaper, less durable and require frequent repairs and replacements [18, 27]. (Painted roofs are relatively uncommon in the study area.) In prior work, roof reflectance and roof color have been shown to be good proxies of housing quality [20, 21]. As such, I aggregated the total building footprint to measure all housing assets (Figure 1e, Building Footprint), and the footprint of tin-roof buildings to measure high-quality housing assets (Figure 1e, Tin-roof Area), in each grid cell. To obtain night light data for systematic comparison, I downloaded and resampled the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) composite images in 2019 [28, 29] (Figure 1e, Night Light).

The maps of treatment intensity and remotely sensed outcomes for the GiveDirectly experiment are shown in Figure 2. For visual display purposes, I plot the maps with a spatial resolution of 0.005° (roughly 500 meters), which is lower than the resolution used in the subsequent statistical analysis. The experiment generated substantial variation in treatment intensity, as expected (Figure 2a). The night light data demonstrate little variation in this rural, sparsely populated area, except in a few spots close to local towns (Figure 2b), whereas both of the housing quality measures capture richer variation in the entire area (Figure 2c, d).

Estimating the Program Effects on Housing Quality. I regressed the remotely sensed outcomes on treatment intensity to estimate the causal effects of the GiveDirectly cash transfer. I chose a spatial resolution of 0.001° (approximately 100m), such that most of the grid cells contain 0–5 households. I exploited only the experimentally-induced random variation in treatment intensity for identification, and accounted for pre-determined differences in program eligibility. Intuitively, imagine two grid cells, one containing a household who received the transfer, and the

other containing a household who was eligible to get the transfer but did not because they were randomized into the control group. With a valid randomization [18], the differences in outcomes between the two can be attributed to the cash transfer. I plot the causal effects on night light and housing quality as cash infusion intensity increases (Figure 3, in red), without making assumptions on the structure of the effects. The results seem to suggest that the effects grow somewhat linearly with the amount of cash infusion. I therefore also report an "average" effect, estimated with the assumption that each $1,000 transfer generated an effect of the same magnitude (Figure 3, panel subtitles). I demonstrate the validity of the empirical strategy further by running 100 placebo simulations—I artificially generated placebo cash transfers that did not actually take place but was consistent with the original randomization design, and estimated their treatment effects (Figure 3, in gray). The resulting estimates are reassuringly centered around zero and not biased.

I observed statistically significant and economically sizable effects on housing quality, on both the extensive margin (larger building footprint) (Figure 3b), and the intensive margin (higher quality roofs) (Figure 3c). On average, a $1,000 cash transfer significantly increased building footprint by 7.9 square meters (85.0 square feet), and tin-roof area by 13.6 square meters (146.4 square feet). These increases indicate that households may have built new structures—either primary residences or auxiliary structures, such as kitchens and sheds, expanded their existing structures, and/or upgraded their thatched roofs to tin roofs, an improvement that people commonly used the transfer for [27]. These estimates are consistent with the results from extensive field surveys, which also documented large increases in housing asset values [18].

On the other hand, I did not observe any program effects on night light (Figure 3a), despite the fact that the cash transfer had large positive impacts on many aspects of the recipient households' living standards—food expenditure, consumer durable spending, asset holding, and housing values [18]. The estimated effect is not statistically significant, small in magnitude, and actually slightly negative. This may be because of low demand for electrification [30], or the poor sensitivity of night light in low-income, rural regions [2].

Recovering the Program Effects on Living Standards with Engel Curves. To take the analysis one step further, I tried to recover the program effects on household living standards with a canonical micro-economic concept, the Engel curve [19, 31–33]. The Engel curve describes the relationship between households' wealth (or other measures of living standards) and their consumption of a particular good, such as food or housing (Figure 4a). For example, someone who lives in a larger house are likely to be wealthier. The slope of the Engel curve represents the ratio between the change in house size and the change in wealth. If we know that someone's house size increased, then we could infer that their wealth also grew—as if they were moving up on the Engel curve. Mathematically, we could divide the change in the house size (the estimates in Figure 3) by the slope of the Engel curve (the "scaling factor" in Figure 4a) to infer the corresponding change in wealth (the "estimated" effects in Figure 4b).

The Engel curves can be derived from any geo-coded consumption and expenditure survey, as long as the surveyed households are—or can be re-weighted to be—representative of the sample in the previous treatment effect estimation step. Notably, the sample does not necessarily have to include any one who has received the treatment, opening up the possibilities of using existing data sources (such as the Living Standards Measurement Study) to estimate Engel curves. In this study, I derived the Engel curves from an endline survey of the GiveDirectly trial participants between May 2016 and June 2017, which included 5,744 households who were eligible for the transfer. Of these households, only those assigned to the control group were used for the estimation. In Figure 4a, I show the relationship between survey-based measures of living standards (x-axis) and remotely sensed night light or housing quality measures (y-axis). The Engel curves are estimated parametrically with a linear regression (dotted lines). The non-parametric fit with LOESS (solid lines) shows only small deviations from the linear regression line, indicating that using a simple linear specification is appropriate. The Engel curves are also roughly monotonically increasing, validating the choice of these variables as wealth proxies.

I scaled the program effect on each remotely sensed outcome to compute the "estimated" effect on household wealth, measured by aggregating the values of a variety of assets. In Figure 4b, I compare these "estimated" effects against the "observed" effects, which were derived from rich endline household survey data and taken from Table 1, Column 1 in the original paper [18]. The estimate based on building footprint (USD 465.7 PPP) is both informative and broadly consistent with the observed effect (USD 555.6 PPP). The estimate based on night light is slightly negative and imprecise, bounding the treatment effect at over USD 1,800 PPP. For reference, the entire GiveDirectly cash transfer is worth USD 1,871 PPP (USD 1,000 nominal), so this estimate hardly provides any new information. The estimate based on tin-roof area is biased. The results are qualitatively similar when I distinguish between housing asset (Supplementary Figure S4) and non-housing asset (Supplementary Figure S5), or when I use annual consumption expenditure as the alternative measure of living standards (Supplementary Figure S6).

The bias in the "estimated" effect based on tin-roof area may be due to the violation of a key assumption, that the Engel curve cannot change directly in response to the treatment. For example, if households participate in a program that directly gives them food, we could no longer look at their food consumption to infer the program effects on living standards, because the relationship between the two will be altered. In this case, only households that lived in thatched-roof houses were eligible for the GiveDirectly transfers. Households' usual consumption patterns of high-quality tin roofs might have been affected by this eligibility criteria. The treatment households owned more tin-roof buildings, compared to control households with the same amount of wealth (Supplementary Figure S7). They may have interpreted this as a "labelled" cash transfer [34]. Roof upgrading may have become a more salient investment because of the targeting criteria used. The lump sum nature of the transfer may have led recipient households to spend more on lumpy purchases than they

normally would. Any of these factors could cause the tin-roof area estimates in Figure 4b to be inflated.

These results highlight the importance of using interpretable proxies when evaluating programs with machine learning predictions. An emerging literature is making great progress in mapping poverty with satellite imagery and machine learning with a high spatial granularity at scale [2–9]. Typically, a machine learning model first learns the mapping between the input satellite images and the ground truth labels of wealth or consumption expenditure, assembled from geo-coded household surveys. Then, the model generates predicted poverty maps for every region in the sample, including those with no survey coverage. The model implicitly combines and executes two tasks: (1) extracting semantically meaningful observations of, say, housing quality, agricultural productivity, or infrastructure, from raw satellite images; and (2) inferring living standards from observing the consumption patterns of these private or public goods (equivalent to the Engel curve analysis in this study). While the flexibility of the machine learning models helps improve predictive performance, the difficulty in interpretation makes it almost impossible to know or constrain what private or public goods are identified and utilized by the model. Since black-box machine learning models utilize as much information as possible from the input satellite images, it is very likely that the Engel curves of at least some of the observed goods will change (similarly to the tin-roof area variable shown in Figure 4b), introducing biases in estimated program effects. In this study, I disentangle the two tasks, so that the first task can be framed as a traditional object detection or segmentation task, allowing me to leverage extensive research in computer science; and the second task becomes more transparent, explicit, and the assumptions testable (for example, with Supplementary Figure S7).

Discussion

In the GiveDirectly experiment, I show that in low-income and rural contexts, housing quality can be a better proxy for economic development than night light. On the other hand, night light works well in middle- or high-income regions, or in highly urbanized areas in low-income countries [14, 15]. Here, I provide additional comparison results in rural Mexico, using the 2010 Population and Housing Census [35] (see Supplementary Materials C for technical details). Housing quality metrics based on daytime imagery (Supplementary Figure S8b, d, f) perform consistently better than night light (Supplementary Figure S8a, c, e), although both demonstrate reasonable sensitivity. As a sanity check, population count 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 (Supplementary Figure S8b). It is only modestly correlated with night light (Supplementary Figure S8a). Night light is less sensitive in smaller, less populated villages, a finding that is consistent with prior work [2]. The asset scores, which are constructed as the first principal component of multiple asset ownership variables, display stronger correlations with average house size (Supplementary

Figure S8d) than with night light (Supplementary Figure S8c). This is not driven entirely by housing assets itself being an important component of wealth. In fact, durable asset holdings, which are not directly visible on satellite images, have a similarly strong correlation with remotely sensed proxies (Supplementary Figure S8e, f).

However, 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 the GiveDirectly study area (and most other 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.

Another fundamental limitation to evaluating programs based on satellite imagery is that the program would have to generate impacts on housing, or other observable variables such as agricultural productivity and infrastructure. This is less plausible for certain programs targeted at addressing other development challenges, or less intensive programs. For example, it seems unlikely that the benefits of vaccination campaigns, or teacher training programs could be observed from space. We also cannot observe the movement of people. Migration rates are very low in the GiveDirectly study area [18], but this could present major challenges if used to study other programs (for example, education-related interventions) that naturally impacts mobility.

Methods

Constructing the Treatment Intensity Map. To construct the treatment intensity map, I utilized data from a baseline census, which was conducted by the authors of the original paper in 2014–2015. The census identified all 65,385 households (roughly 280,000 people) residing in 653 villages in the study area, recorded their GPS coordinates, whether each household was eligible for the GiveDirectly cash transfer, and whether they had been randomized into the treatment or control group [18]. To address the measurement errors of the GPS collection devices, I discarded 58 outliers (living more than 2 kilometers away from the village centers) and imputed those and other 4 missing GPS coordinates with village center coordinates. Then, I converted these household records into a raster map. I laid out a regular grid, and counted, in each grid cell, the number of households that ultimately received the GiveDirectly cash transfer (see Figure 1 and Figure 2a). Grid cells containing no eligible households were excluded. To account for pre-determined policy intensity differences, I recorded (and later controlled for) the number of households that were eligible for the cash transfer, regardless of whether they had been randomized into the treatment or control group.

Obtaining High-resolution Daytime Satellite Images. I utilized high-resolution daytime satellite images from Google Static Maps [24]. These images have a spatial resolution of about 30cm per pixel (at equator), and contain only the RGB (red, green, blue) bands (see Figure 1c and Supplementary Figure S2 for examples). These images come from a variety of commercial providers such as Maxar (formerly DigitalGlobe) and Airbus, and have been seamlessly mosaicked together. They have also been geo-referenced and pre-processed to remove clouds and address other data quality issues. Google does not provide the exact timestamps for these images, but I estimate that they were taken in 2019, most likely on Dec 30, 2019. The dates for retrieving these images from the Google Static Maps API are between Feb 19 and Feb 21, 2020, and the Google Earth Pro imagery archive reflects that the closest available images in the study area were from Dec 30, 2019. Multiple other satellite images taken in February, March, July, August and September 2019 are also available in the study area, indicating that the images used in this study are most certainly from 2019.

Extracting Housing Quality Metrics with Mask R-CNN. I first leveraged a state-of-the-art deep learning model, Mask R-CNN [17] to segment buildings—that is, to detect each building and the pixels that they occupy—in the Google Static Map satellite images. I then converted the pixel-wise predictions to polygons, and extracted housing quality metrics related to the size of the building and the roof materials from each polygon (see Figure 1d and Supplementary Figure S2 for examples).

Loosely speaking, the Mask R-CNN model operates as follows. First, the model proposes a large number of "regions of interest", each of which potentially contains a building. Then, the

model uses convolutional filters to identify patterns within the proposed region that are indicative of the presence of buildings, such as the sharp edges, the highly reflective roofs, and the building shadows. Finally, the model predicts whether each proposed region contains a building, as well as whether each pixel is occupied by the building.

I trained the Mask R-CNN model with a multi-step process and a transfer learning framework, as described in greater detail in Supplementary Materials B. Publicly available building footprint datasets in rural and low-income regions are rare, and they often differ substantially in spatial resolution, sensor instrument, and landscape from inference images (that is, the target images that the model will make predictions for). Relying solely on publicly available training data is therefore insufficient for achieving satisfactory predictive performance. I curated a set of in-sample annotations by randomly sampling 120 images from all the Google Static Map images in the study area, and manually creating high-quality building footprint annotations for them. I pre-trained the Mask R-CNN model on large, publicly available datasets such as COCO (Common Objects in Context) and Open AI Tanzania, and fine-tuned them on this set of in-sample annotations.

The model predictions are highly accurate. The overall F1 score (a standard performance metric for instance segmentation) on a random subset of inference images is 0.79 (Supplementary Figure S1). The F1 score is the harmonic mean of precision (the proportion of model-identified buildings that are actual buildings) and recall (the proportion of actual buildings that are correctly identified by the model). Here, a building is deemed to be correctly identified if the predicted pixel mask and the ground truth pixel mask have sufficient overlap (more precisely, if the intersection of the two masks is more than 50% of the union of the two masks). As a reference point, the top winner in the 2nd SpaceNet building footprint extraction competition reported an F1 score of 0.69 [36]. This demonstrates that the Mask R-CNN model used in this study performs well, although building footprint segmentation in rural, less complex scenes is generally easier than in modern cities so these metrics are not directly comparable.

I post-processed the model-predicted pixel masks by converting them to polygons, and simplifying the polygons with the Douglas-Peucker algorithm with a pixel tolerance of 3. For each polygon, I computed two housing quality metrics: building footprint and type of roof materials. I then laid out a regular grid, assigned each building to grid cells based on the centroids of the polygons, and aggregated to obtain two metrics at the pixel level: building footprint (Figure 2c) and tin-roof area (Figure 2d).

First, I measured the size of each building polygon and converted it to square meters. I corrected for area distortion, which is induced by the Web Mercator projection system that the Google Static Map uses. This metric may appear larger than what one expects for the size of homes in a low-income context (Figure 4), because (1) it represents the footprint of the entire building, which is typically larger than the size of the livable area; and (2) it accounts for both residential and non-residential structures, since the model is not able to distinguish between the two.

Second, I estimated the types of roof materials based on the colors of the roofs, and computed the footprint of tin-roof buildings in each grid cell. For each building, I took all the pixels associated with the given building instance, and assigned a "representative" roof color by computing the average values in the RGB (Red, Green, Blue) channels. Since the Euclidean distances between color vectors in the RGB color space does not reflect perceptual differences, I projected all the RGB color vectors to the CIELAB color space, and clustered these roof color vectors into 8 groups by running the K-means clustering algorithm. I further classified these 8 groups into three types of roof materials: tin roof, thatched roof, and painted roof (Supplementary Figure S3), and computed the total footprint of tin-roof buildings.

Obtaining the Night Light Data. To measure nighttime luminosity, I used the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) composite images hosted on Google Earth Engine [28, 29]. The VIIRS-DNB data product excludes areas impacted by cloud cover and correct for stray light [37]. However, it has not been filtered to screen out lights from aurora, fires, boats, and other temporal lights, and lights are not separated from background (non-light) values [28]. This data product has a native spatial resolution of 15 arc seconds (approximately 463 meters at the equator), and I resampled the data by conducting nearest neighbor interpolation when necessary. I averaged over all the monthly observations in 2019 and constructed a single cross sectional observation, to reduce seasonality effects and for consistency with the daytime satellite imagery (Figure 2b). The VIIRS-DNB data product is considered superior to the more widely used night light data, DMSP-OLS (the United States Air Force Defense Meteorological Satellite Program, Operational Linescan System) because it preserves finer spatial details, has a lower detection limit and displays no saturation on bright lights [38]. This ensures that I conduct a fair comparison with the most modern and high-quality night light data product.

Estimating the Program Effects on Housing Quality. The main econometric specification for Figure 3 is as follows

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

where each observation i represents a $0.001^\circ \times 0.001^\circ$ grid cell (approximately 100m$\times$100m); τ_k represents the estimate of interest: the treatment effects of the unconditional cash transfer on remotely sensed outcomes; x_i denotes the number of recipient households per grid cell (equivalent to the amount of cash infusion in \$1000); e_i denotes the number of eligible households per grid cell, with $m \in M = \{0, 1, 2, 3, \cdots\}$; and y_i denotes remotely sensed outcomes: night light, building footprint, and tin-roof area. To account for pre-existing differences in population density or wealth, which may cause non-random variation in treatment intensity, I flexibly controlled for the number of eligible households per grid cell, and excluded grid cells with no eligible households. Because the

grid cells are fairly small and the number of observations for $k > 2$ is small, I binned the number of recipient households into four bins $k \in K = \{0, 1, 2, 2+\}$, to preserve statistical power. Standard errors were calculated à la Conley, with a uniform kernel and a 3km cutoff [39–42]. To reduce the effects of outliers (due to sensor malfunctioning or machine learning model prediction errors), I winsorized all remotely sensed variables at the 97.5 percentile.

I ran 100 placebo simulations to further demonstrate the validity of the main specification. In each simulation, I randomly assigned half of the 68 groups of villages to the high-saturation group, and the other half to the low-saturation group. In the high-saturation groups, I randomly assigned 2/3 of the villages to the treatment group (and the rest to the control group); whereas in the low-saturation group, I assigned only 1/3 of the villages to the treatment group (and the rest to the control group). This mimics the two-tier randomization scheme of the original trial [18]. Using these simulated placebo treatment status variables, I estimated the placebo treatment effects with the econometric specification described in Equation 1.

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

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

where τ is the "average" treatment effect, and all else remain the same as in Equation 1.

Estimating the Engel Curves. An Engel curve describes how household expenditure on a particular good varies with income—a relationship that can be used to infer households' living standards from the consumption patterns of a limited subset of goods [19, 31–33]. The mathematical formulation is

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

where household h with W_h wealth (or other measures of living standards) would consume Q_{hp} quantities of a normal good p, and $F_p(\cdot)$ represents the Engel curve for product p in the population. With a linearity assumption, this can be simplified to be

$$Q_{hp} = \alpha_p + \beta_p W_h + \epsilon_{hp} \tag{4}$$

where α_p is the intercept and β_p is the slope of a linear Engel curve.

In this study, I estimated the Engel curves based on the endline survey of the original GiveDirectly trial, which included a representative set of 5,744 households who were eligible for the transfer. The households participated in a comprehensive consumption and expenditure survey between May

2016 and June 2017, after the distribution of cash transfers. From the surveys, I observed annualized household consumption expenditure, and asset values. 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. Household assets include housing and non-housing assets, but not land values. Housing asset values are measured as the respondent's self-reported cost to build a home like theirs. Non-housing assets include livestock, transportation (bicycles, motorcycles, and cars), electronics, farm tools, furniture, other home goods, and lending or borrowing from formal or informal sources. I did not study land values because they are difficult to value given thin local markets [18].

I performed heuristic matching between the buildings and the household survey GPS coordinates, to link variables in the survey with remotely sensed variables. First, I took the baseline census data, which geo-coded every single household who lived in the study area, and assigned every building in the satellite images to its closest census GPS coordinate, if the distance between the two was within 250m. This ensures that every building is matched to at most one household. Second, I matched GPS coordinates from the survey with GPS coordinates from the census. While the same household supposedly had the same geo-location, these two often differed because of the measurement errors of the GPS collection devices, and because the coordinates might be recorded anywhere on the participants' plots and not necessarily in their primary residence. I similarly assigned each survey GPS coordinate to its closest census GPS coordinate, if the distance between the two was within 250m. In cases of multiple surveys being assigned to the same census coordinate, I kept the closest survey. The final sample contains only census observations that are matched with both buildings in the satellite images and survey records, and consists of 4,594 households.

The Engel curves were estimated with only the control group (Figure 4a and Supplementary Figure S4a, S5a and S6a). They were estimated both non-parametrically with LOESS (see Equation 3 and the solid lines in Figure 4a) and parametrically with a linear regression (see Equation 4 and the dotted lines in Figure 4a). When fitting LOESS, I allowed for locally-fitted quadratic polynomials, and used 75% of the data points for each fit. To minimize the influence of outliers, I winsorized consumption expenditure, housing asset values, non-housing asset values and overall asset values at the 97.5 percentile of the full (eligible and non-eligible) sample. Additionally, I winsorized non-housing asset values and overall asset values at the 2.5 percentile, as outliers with a large amount of debt exist and could potentially drive the results otherwise. I also winsorized all the remotely sensed variables at the 97.5 percentile.

Recovering the Program Effects on Living Standards. I adapted a prior mathematical formulation that used the Engel curve to infer changes in living standards [19]. Suppose that one is interested in studying the effect of a plausibly exogenous treatment Z on, say, wealth W (denoted $\hat{\tau}_W$), but can only inexpensively observe its effect on the consumption of product p (denoted $\hat{\tau}_{Q_p}$).

Recall that $\hat{\beta}_p$ is the estimated slope of the linear Engel curve in Equation 4, then

$$\hat{\tau}_W = \hat{\tau}_{Q_p} / \hat{\beta}_p \tag{5}$$

Using a formula for propagation of error (or the multivariate Delta method), one can derive the standard error for $\hat{\tau}_W$ as follows. This derivation is based on prior work [19], but additionally accounts for the precision of the slope of the Engel curve.

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

A key assumption of this approach is that $\hat{\beta}_p$ does not depend on Z—that is, the Engel curve does not change in direct response to the treatment—also termed the conditional independence assumption [32].

I computed the "estimated" treatment effects on wealth (or other living standard measures) according to Equation 5 and Equation 6, with the treatment effect estimates for remotely sensed variables, and the slopes of the Engel curves. I compared these "estimated" effects against the "observed" effects, taken from Table 1, Column 1 in the original paper [18], which were estimated based on the endline household survey data (Figure 4b).

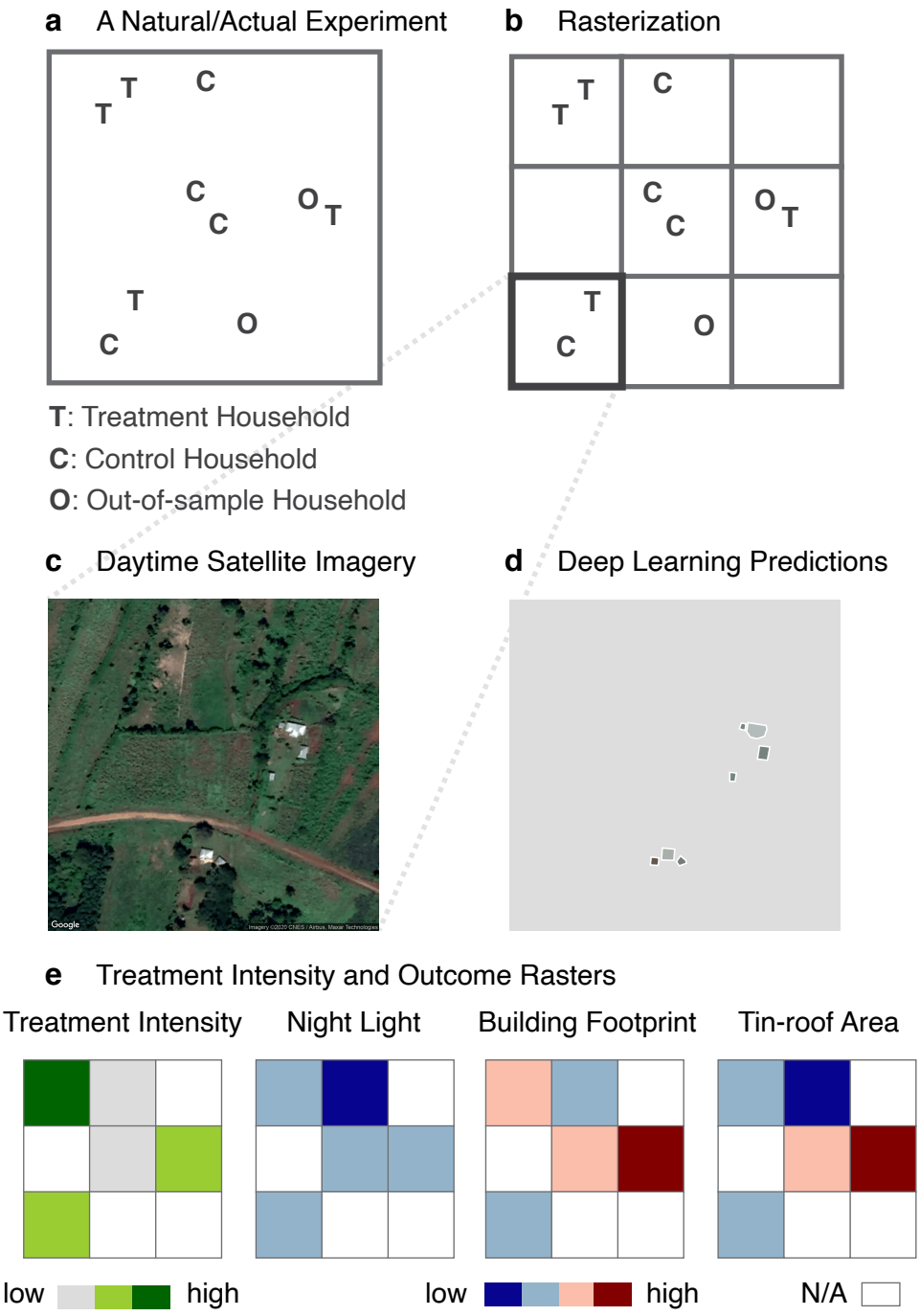

Figure 1: **Constructing maps of treatment intensity and remotely sensed outcomes from program implementation records and satellite imagery.** **a** An illustration of geocoded program implementation records. **b** Placing a regular grid over **a** and measuring the intensity of the treatment in each grid cell. **c** An example daytime satellite image from Google Static Maps. **d** Example deep learning predictions on **c**. Each building is outlined in white and filled with the "representative" roof color. **e** The rasters of treatment intensity (constructed from **b**) and remotely sensed outcomes (constructed from **d**). Grid cells without in-sample households are omitted.

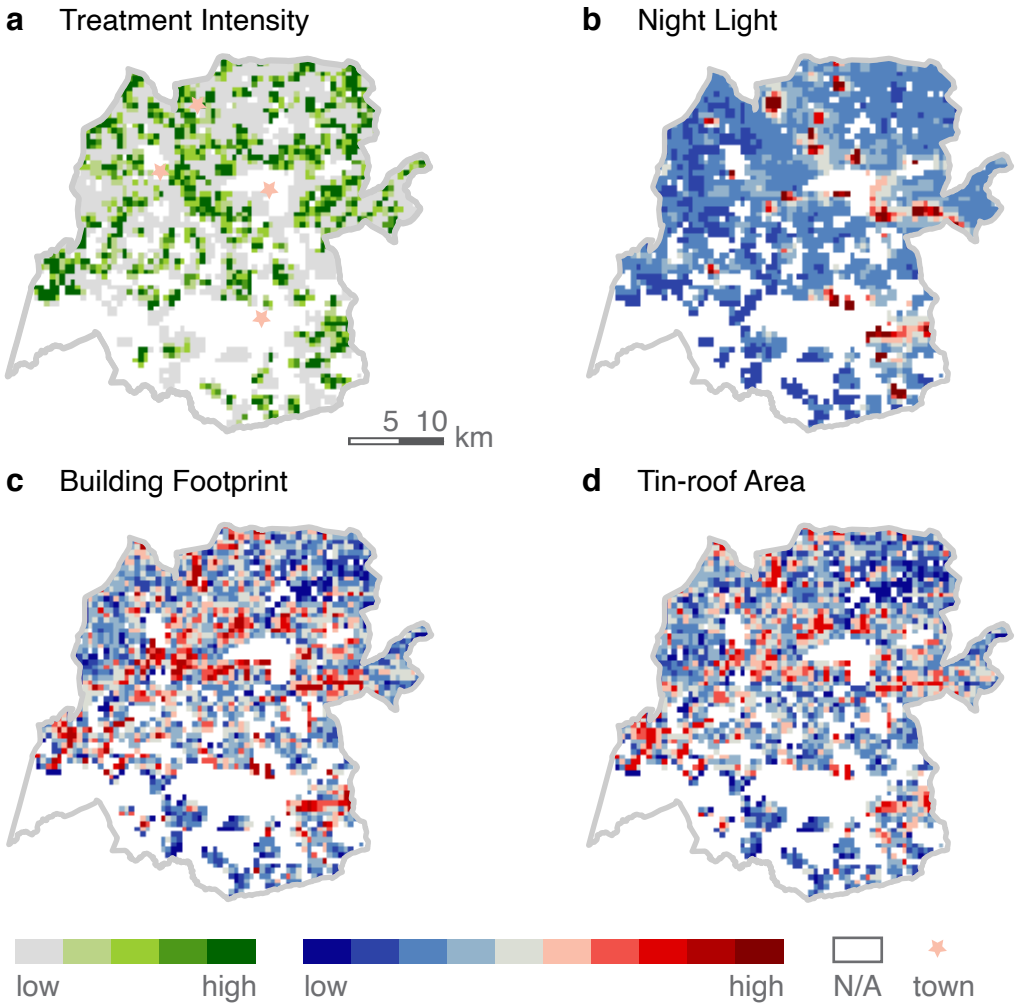

Figure 2: **Mapping treatment intensity and remotely sensed outcomes in the GiveDirectly study area in 2019.** **a** Treatment intensity represents the number of households who received a $1,000 cash transfer from GiveDirectly. **b** Night light is the average radiance in the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB). **c** Building footprint measures the total area covered by any building. **d** Tin-roof area measures the total footprint of buildings with roofs made of tin, a high quality construction material. In all the panels, the gray lines outline the GiveDirectly study area in Siaya, Kenya. Grid cells without in-sample households are omitted.

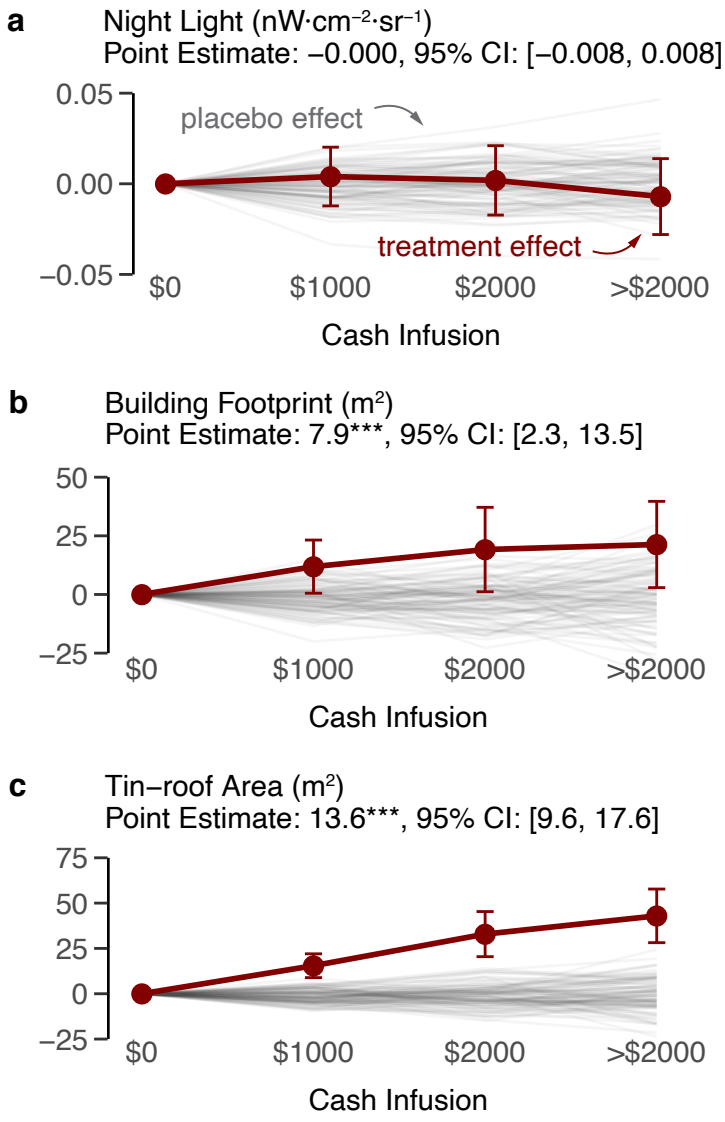

Figure 3: **Housing quality increased in response to the GiveDirectly cash transfer, but night light remained unchanged.** The treatment effects of the cash transfers on night light (**a**), building footprint (**b**), and tin-roof area (**c**) are shown in red. The dots represent the point estimates, and the error bars represent the 95% confidence intervals. Gray lines show the estimated effects of the placebo cash infusions from 100 simulations. The average treatment effects of a $1,000 transfer (and their 95% confidence intervals), estimated based on a constant effect assumption, are reported in the panel subtitles. *** indicates statistical significance at the 1% level.

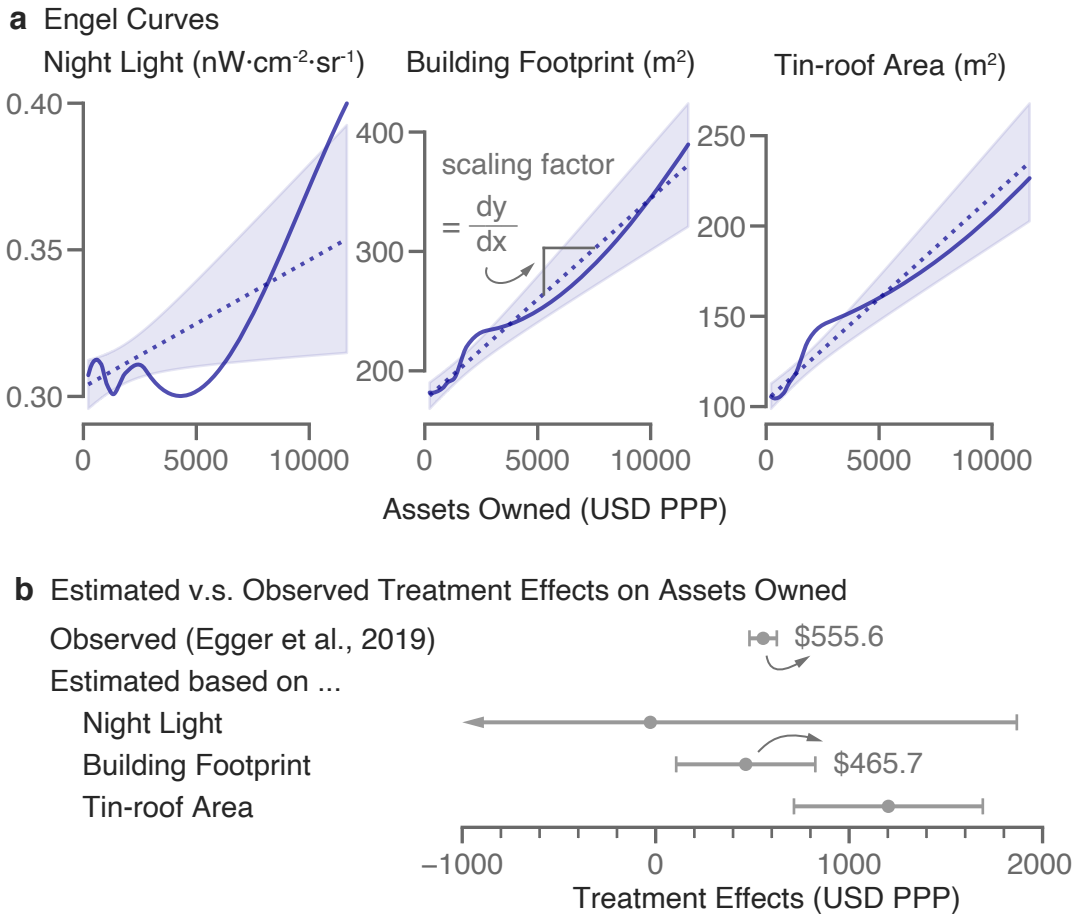

Figure 4: **The treatment effect on household assets can be correctly recovered by scaling the effect on building footprint. a** The Engel curves of night light, building footprint, and tin-roof area, estimated non-parametrically (solid line) or parametrically (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the estimated versus observed treatment effects. The dot shows the point estimate. The error bars show the 95% confidence intervals, with the arrow marking numbers that are out of range.

References

1. *Understanding GiveDirectly's finances* https://www.givedirectly.org/financials/. accessed 4 Feb 2021.
2. Jean, N. *et al.* Combining satellite imagery and machine learning to predict poverty. *Science* **353,** 790–794 (2016).
3. Blumenstock, J. E. Fighting poverty with data. *Science* **353,** 753–754 (2016).
4. Yeh, C. *et al.* Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications* **11,** 1–11 (2020).
5. Aiken, E. L., Bedoya, G., Coville, A. & Blumenstock, J. E. *Targeting Development Aid with Machine Learning and Mobile Phone Data: Evidence from an Anti-Poverty Intervention in Afghanistan* in *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies* (Association for Computing Machinery, Ecuador, 2020), 310–311. https://doi.org/10.1145/3378393.3402274.
6. Blumenstock, J. Machine learning can help get COVID-19 aid to those who need it most. *Nature* (2020).
7. Watmough, G. R. *et al.* Socioecologically informed use of remote sensing data to predict rural household poverty. *Proceedings of the National Academy of Sciences* **116,** 1213–1218 (2019).
8. Engstrom, R., Hersh, J. & Newhouse, D. *Poverty from space: using high-resolution satellite imagery for estimating economic well-being* tech. rep. (The World Bank, 2017).
9. Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A. & Swartz, T. *Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images, With an Application in Mexico* in *Proceedings of NIPS 2017 Workshop on Machine Learning for the Developing World* (2017). https://arxiv.org/abs/1711.06323.
10. Deaton, A. *The analysis of household surveys: a microeconometric approach to development policy* (The World Bank, 1997).
11. Banerjee, A. V. & Duflo, E. *Poor economics: A radical rethinking of the way to fight global poverty* (Public Affairs, 2011).
12. Pamies-Sumner, S. *Development Impact Evaluations: State of Play and New Challenges* tech. rep. (Agence Française de Développement, 2015).

13. Brune, L., Karlan, D., Kurdi, S. & Udry, C. R. *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).

14. Henderson, J. V., Storeygard, A. & Weil, D. N.
Measuring economic growth from outer space. *American economic review* **102,** 994–1028 (2012).

15. Chen, X. & Nordhaus, W. D. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences* **108,** 8589–8594 (2011).

16. Michalopoulos, S. & Papaioannou, E.
National institutions and subnational development in Africa.
The Quarterly journal of economics **129,** 151–213 (2014).

17. He, K., Gkioxari, G., Dollár, P. & Girshick, R. *Mask R-CNN*
in *Proceedings of the IEEE international conference on computer vision* (2017), 2961–2969.

18. Egger, D., Haushofer, J., Miguel, E., Niehaus, P. & Walker, M. W.
General equilibrium effects of cash transfers: experimental evidence from Kenya tech. rep. (National Bureau of Economic Research, 2019).

19. Young, A. The African growth miracle. *Journal of Political Economy* **120,** 696–739 (2012).

20. Marx, B., Stoker, T. M. & Suri, T.
There is no free house: Ethnic patronage in a Kenyan slum.
American Economic Journal: Applied Economics **11,** 36–70 (2019).

21. Michaels, G. *et al.*
Planning Ahead for Better Neighborhoods: Long Run Evidence from Tanzania tech. rep. (IZA Institute of Labor Economics, 2017).

22. Kohler, T. A. *et al.* Greater post-Neolithic wealth disparities in Eurasia than in North America and Mesoamerica. *Nature* **551,** 619–622 (2017).

23. OECD. in *National Accounts at a Glance 2014* (OECD Publishing, Paris, 2014).

24. *Google Static Maps* https://developers.google.com/maps/documentation/maps-static/intro. accessed 6 May 2020.

25. *COCO - Common Objects in Contexts* http://cocodataset.org. accessed 6 May 2020.

26. *2018 Open AI Tanzania Building Footprint Segmentation Challenge*
https://competitions.codalab.org/competitions/20100. accessed 6 May 2020.

27. Haushofer, J. & Shapiro, J. The short-term impact of unconditional cash transfers to the poor: experimental evidence from Kenya.
The Quarterly Journal of Economics **131,** 1973–2042 (2016).

28. *VIIRS Stray Light Corrected Nighttime Day/Night Band Composites Version 1* https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_MONTHLY_V1_VCMSLCFG. accessed 6 May 2020.

29. Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C. & Ghosh, T. VIIRS night-time lights. *International Journal of Remote Sensing* **38,** 5860–5879 (2017).

30. Lee, K., Miguel, E. & Wolfram, C. Experimental evidence on the economics of rural electrification. *Journal of Political Economy* **128,** 1523–1565 (2020).

31. Elbers, C., Lanjouw, J. O. & Lanjouw, P. Micro–level estimation of poverty and inequality. *Econometrica* **71,** 355–364 (2003).

32. Tarozzi, A. & Deaton, A. Using census and survey data to estimate poverty and inequality for small areas. *The review of economics and statistics* **91,** 773–792 (2009).

33. Atkin, D., Faber, B., Fally, T. & Gonzalez-Navarro, M. *A New Engel on Price Index and Welfare Estimation* tech. rep. (National Bureau of Economic Research, 2020).

34. Benhassine, N., Devoto, F., Duflo, E., Dupas, P. & Pouliquen, V. Turning a shove into a nudge? A "labeled cash transfer" for education. *American Economic Journal: Economic Policy* **7,** 86–125 (2015).

35. *2010 Population and Housing Census of Mexico* https://www.inegi.org.mx/programas/ccpv/2010/default.html. accessed 5 May 2020.

36. *2nd SpaceNet Competition Winners Code Release* https://medium.com/the-downlinq/2nd-spacenet-competition-winners-code-release-c7473eea7c11. accessed 29 Jan 2021.

37. Mills, S., Weiss, S. & Liang, C. *VIIRS day/night band (DNB) stray light characterization and correction* in *Earth Observing Systems XVIII* **8866** (2013), 88661P.

38. Elvidge, C. D., Baugh, K. E., Zhizhin, M. & Hsu, F.-C. Why VIIRS data are superior to DMSP for mapping nighttime lights. *Proceedings of the Asia-Pacific Advanced Network* **35** (2013).

39. Conley, T. G. GMM estimation with cross sectional dependence. *Journal of econometrics* **92,** 1–45 (1999).

40. Conley, T.
Spatial econometrics. New Palgrave Dictionary of Economics, eds Durlauf SN, Blume LE 2008.

41. Hsiang, S. M. 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).

42. Burlig, F. & Woerman, M. *ARE 212 Section 10: Non-Standard Standard Errors II*
https://static1.squarespace.com/static/558eff8ce4b023b6b855320a/t/573bd63745bf21da74c080a8/1463539276997/ARE_212_Section_10.pdf.
accessed 10 May 2020.

Supplementary Materials

A Supplementary Figures

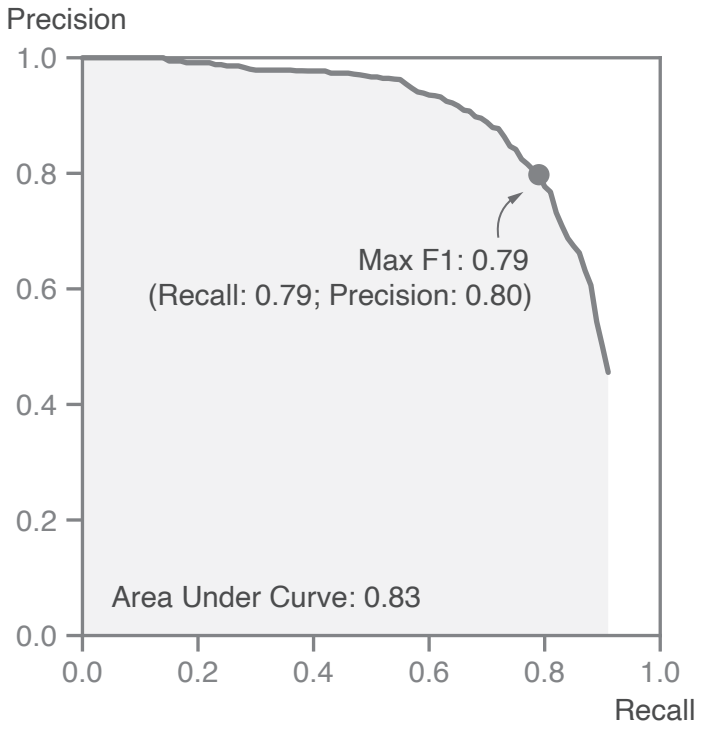

Figure S1: **The precision-recall curve of the Mask R-CNN model shows satisfactory predictive performance.** The Mask R-CNN model is trained and evaluated with 3-fold cross validation. The evaluation is based on 120 annotated images, which were randomly sampled from all the input satellite images in Siaya, Kenya. The Mask R-CNN model outputs a confidence score for every predicted building instance, and the precision-recall curve is generated by varying the confidence score threshold, below which predicted instances are dropped. A higher threshold makes the model more conservative and corresponds to the left portion of the curve (with high precision and low recall), and vice versa. The dot represents the optimal confidence score threshold, obtained by maximizing F1, the harmonic mean of precision and recall. The main model used in this study employs the optimal threshold, and has a recall of 0.79 and a precision of 0.80.

Figure S2: **Ten randomly sampled pairs of input images and deep learning predictions.** Ten images are randomly sampled from all the input satellite images in the GiveDirectly study area. Each predicted building is outlined in white and filled with the "representative" roof color.

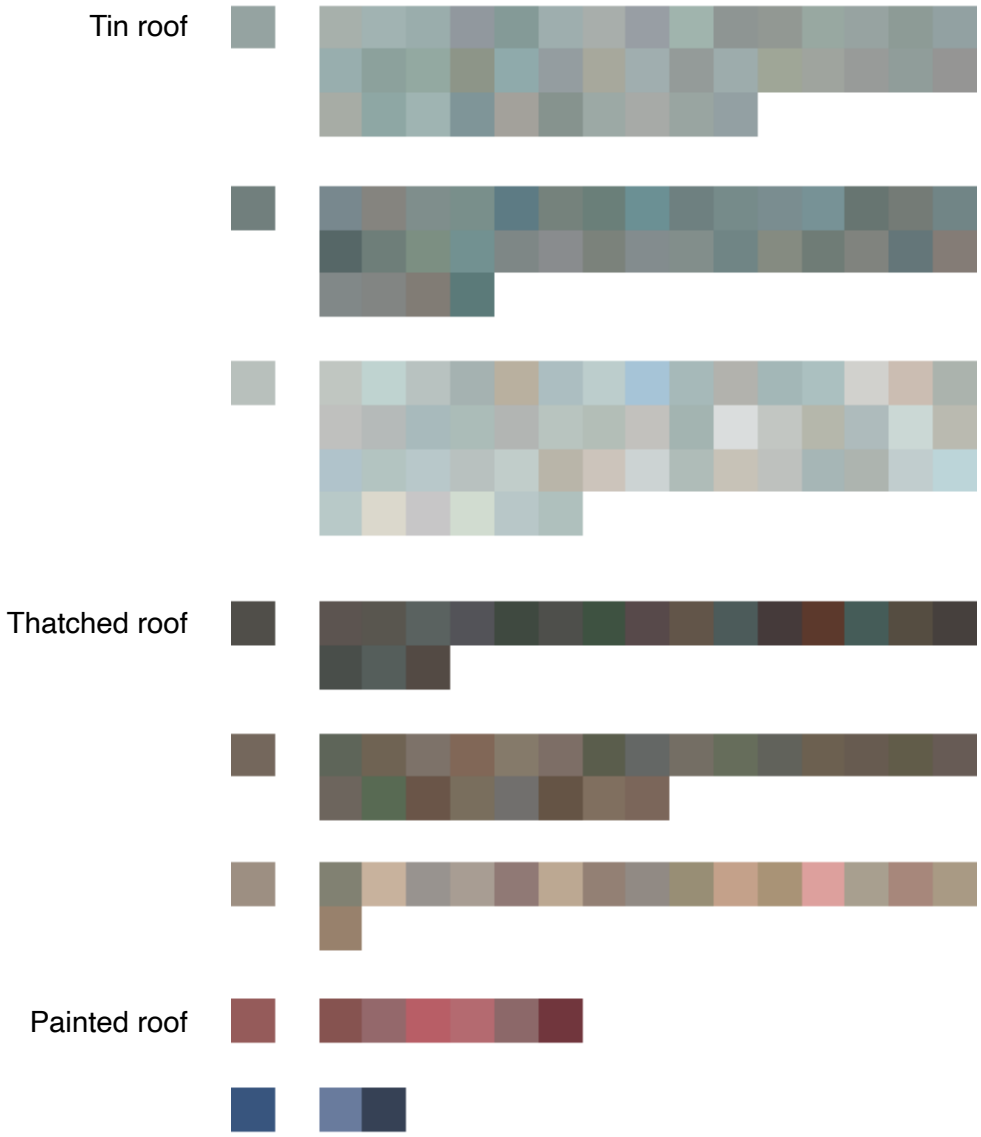

Figure S3: **The distribution and grouping of roof colors.** All the buildings in the GiveDirectly study area are split into eight groups by a K-means clustering algorithm, based on their roof colors. The color block on the left represents the "average" roof color of the cluster, and the color blocks on the right represent a random subset of all the roof colors in the given cluster. The number of color blocks on the right is proportional to the size of the cluster. The eight groups are further grouped into tin roof, thatched roof, and painted roof.

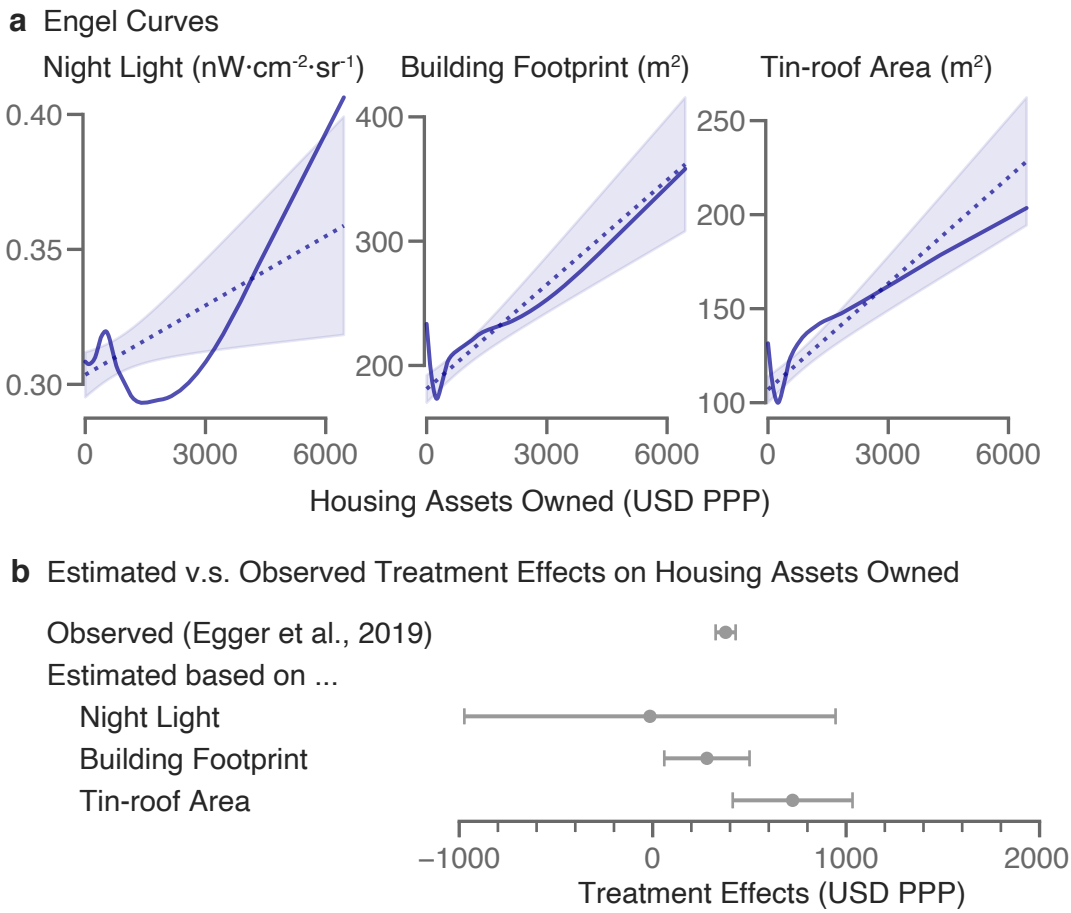

Figure S4: **The treatment effect on housing assets can be similarly recovered by scaling the effect on building footprint.** **a** The Engel curves of night light, building footprint, and tin-roof area, estimated non-parametrically (solid line) or parametrically (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the estimated versus observed treatment effects. The dot shows the point estimate. The error bars show the 95% confidence intervals.

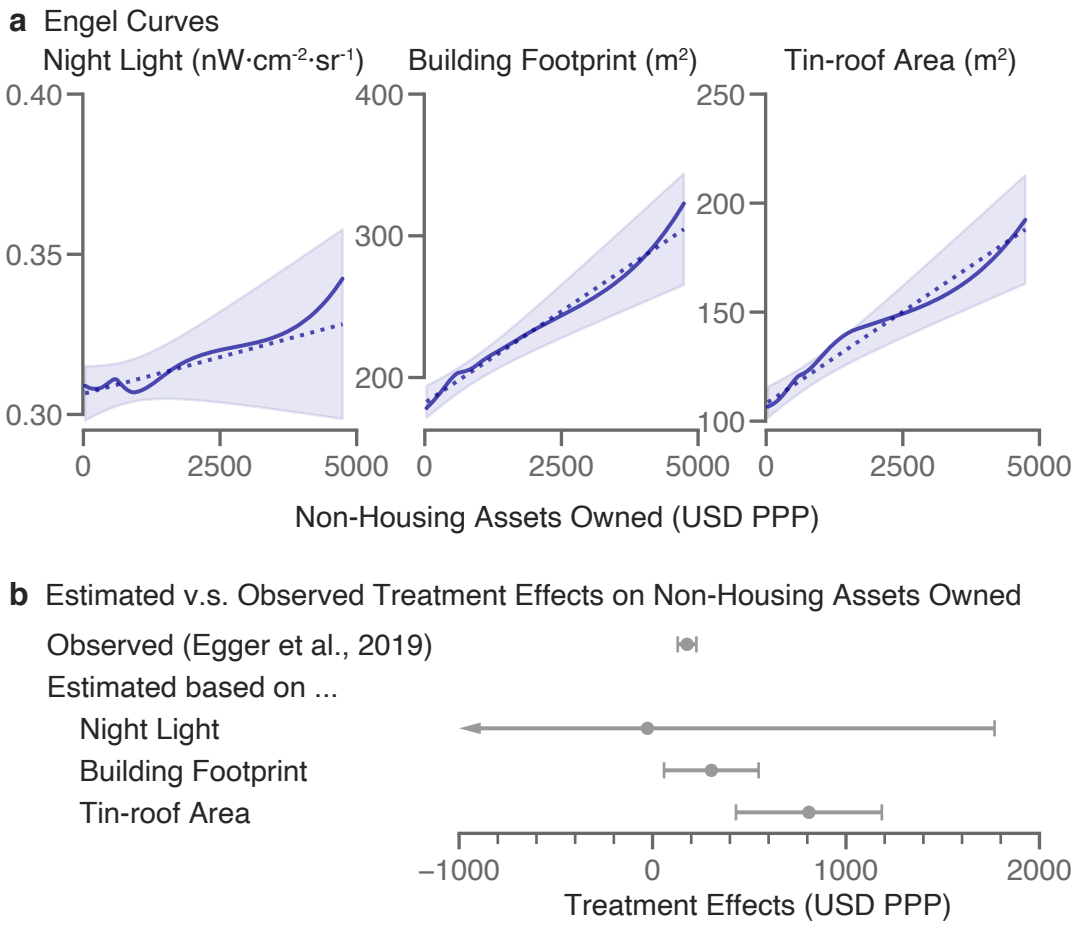

Figure S5: **The treatment effect on non-housing assets can be similarly recovered by scaling the effect on building footprint. a** The Engel curves of night light, building footprint, and tin-roof area, estimated non-parametrically (solid line) or parametrically (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the estimated versus observed treatment effects. The dot shows the point estimate. The error bars show the 95% confidence intervals, with the arrow marking numbers that are out of range.

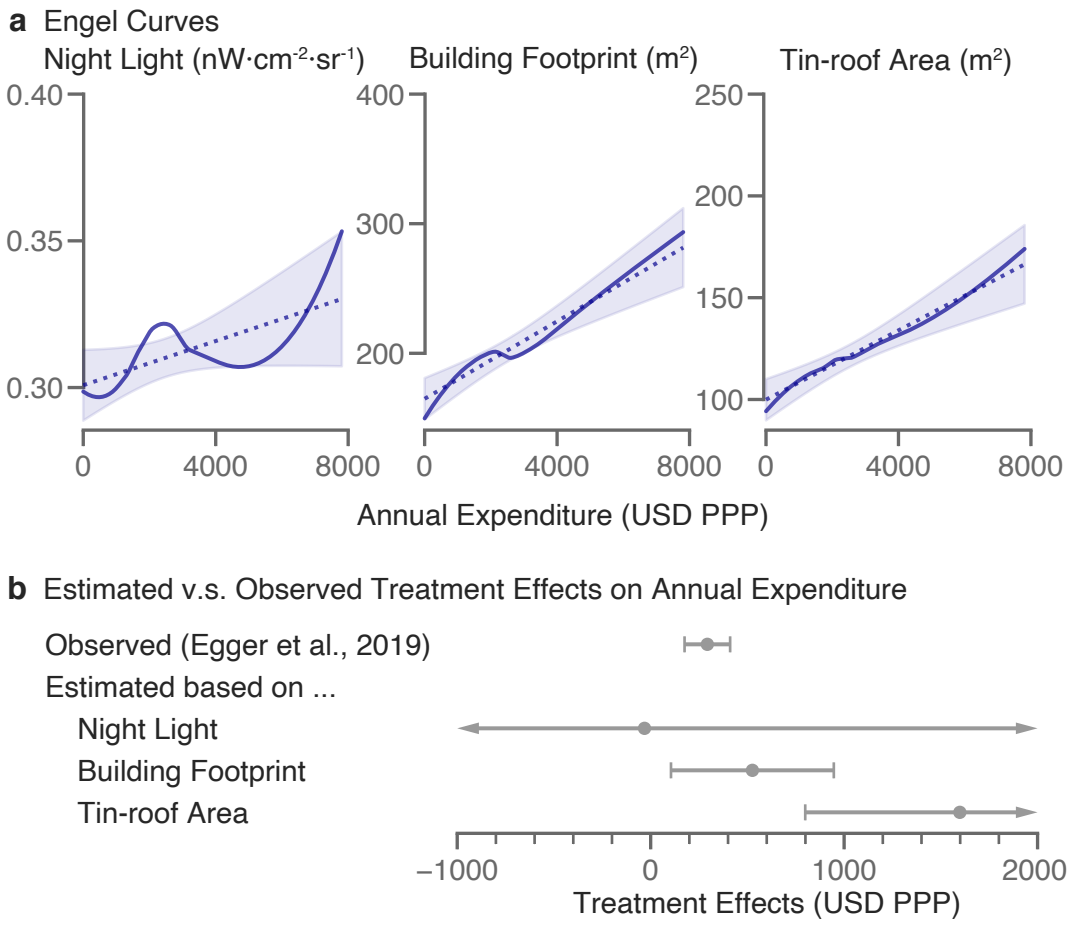

Figure S6: **The treatment effect on annual consumption expenditure can be similarly recovered by scaling the effect on building footprint.** **a** The Engel curves of night light, building footprint, and tin-roof area, estimated non-parametrically (solid line) or parametrically (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the estimated versus observed treatment effects. The dot shows the point estimate. The error bars show the 95% confidence intervals, with the arrow marking numbers that are out of range.

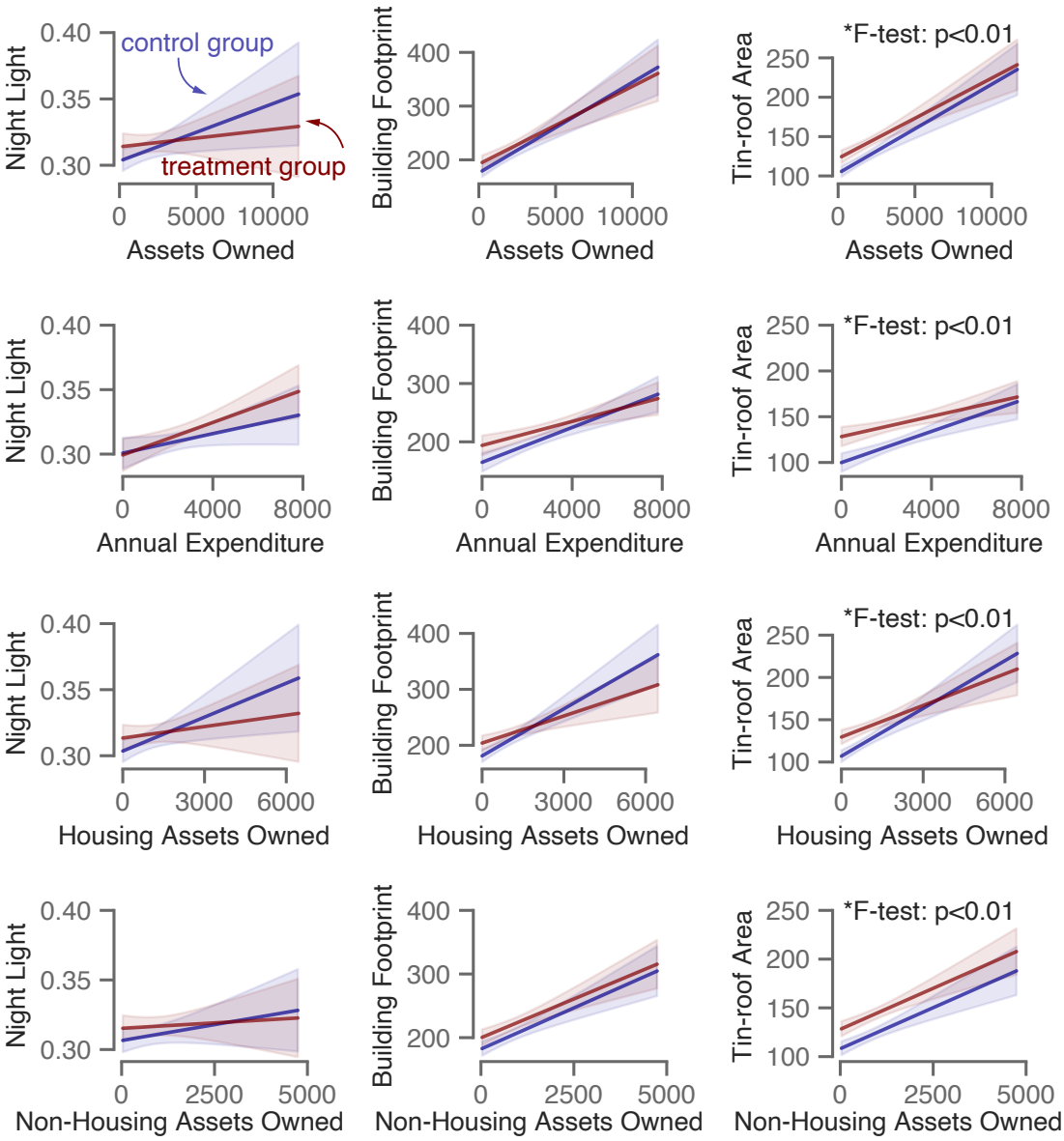

Figure S7: **The Engel curves for tin-roof area shifted in response to the cash transfer.** The Engel curves for the treatment households (in red) and the control households (in blue). The shaded regions represent the 95% confidence intervals. The F-test tests for differences in the slopes and intercepts of the regression lines between the treatment and the control households. p values that are statistically significant at the 1% level are marked.

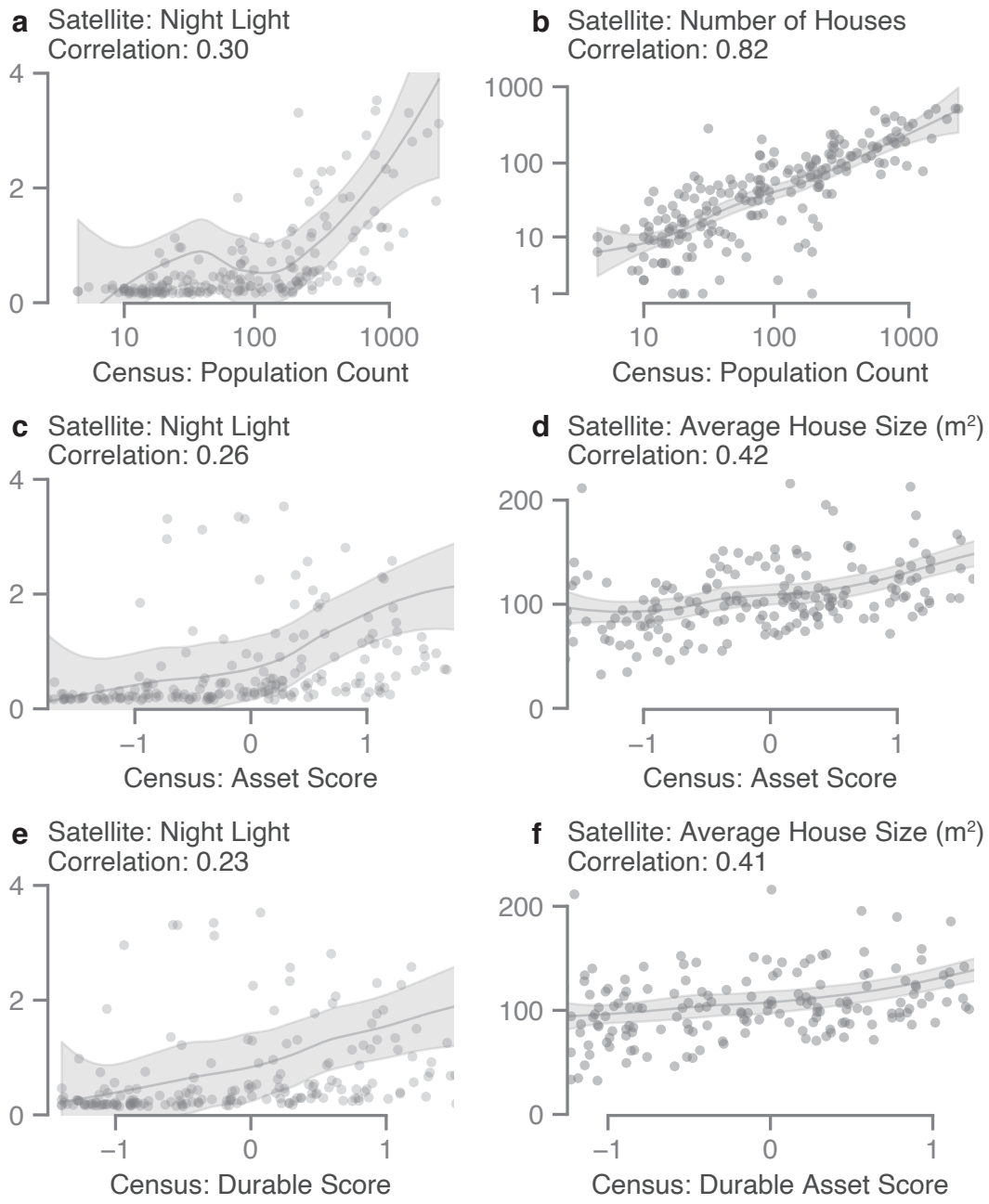

Figure S8: **Remotely sensed housing quality measures correlate with population count and household wealth more strongly than night light in rural Mexico.** The relationships between remotely sensed variables (y axis) and census-derived variables (x axis). The population counts are shown in log scale. Each point corresponds to a randomly sampled rural locality in Mexico. Gray lines are estimated LOESS curves, and the shaded regions are the 95% confidence intervals. The (Pearson) correlation coefficients are reported in the panel subtitles.

B Training the Deep Learning Model

B.1 Creating In-sample Building Footprint Annotations

I created in-sample building footprint annotations to train the model, and to objectively and quantitatively evaluate model performance. Among the 71,012 satellite images that cover all of the Siaya county in Kenya, I randomly sampled 120 images for annotation. I used the Supervisely image annotation web platform to create annotations. On any given image, I outlined the boundaries of all the instances of buildings on the image. Buildings that border each other were annotated as separate instances, if there were reasons to believe that they were separate structures (e.g., if they appeared to use different roof materials). Half-finished buildings were annotated, although they were fairly rare in the analysis sample.

Some measurement errors can arise from the annotation process, which may in turn impact the predictions of the deep learning model. First, the Google Static Maps logo blocks 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, the annotation accuracy (and thus potentially prediction accuracy) may be different across buildings with different roof materials. In particular, thatched-roof houses tend to be harder to identify for human annotators than metal-roof houses, because they are typically smaller, not as reflective, and may resemble trees in the overhead imagery.

B.2 Training the Mask R-CNN Model

I used the Mask R-CNN model [1] for instance segmentation of buildings on satellite images. The backbone architecture used was ResNet50 with the Feature Pyramid Networks. The model was trained with a learning rate of 5×10^{-4} and a batch size of 10. Optimization was conducted with the Adam optimizer. I implemented the deep learning pipeline with Python and PyTorch. In particular, I used the official Torchvision implementation of Mask R-CNN. I trained the Mask R-CNN model in a transfer learning framework, with a multi-step process as follows.

1. COCO (Common Objects in Context) The model was 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 [2]. 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.

2. Open AI Tanzania The model was then 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 [3]. These images are 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. All the buildings in the drone images were 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 are a large number of unfinished structures in Zanzibar. Most input satellite images in this study contain very few unfinished structures, so I collapsed the first two categories into one and drop the third category. The native resolution of the drone images is 7cm, and I down-sampled the images to about 30cm to match with the resolution of the target satellite images.

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

3. Supplementary Annotations in Mexico, Tanzania and Kenya The model was then fine-tuned on a set of 587 annotated high-resolution satellite images from Mexico, Tanzania, and Kenya. The Mexico dataset consists of 199 satellite images corresponding to 8 randomly sampled rural localities studied in Supplementary Figure S8. Some of these are historical images with lower data quality and more cloud coverage. These images were pooled and randomly split into a training set (90%) and a validation set (10%). The model was trained for 25 epochs, and achieved the best performance at epoch 17.

4. In-sample Annotations Finally, the model was fine-tuned on a set of 120 in-sample annotated images in Siaya, Kenya (see Section B.1 for details). This ensures that training images and inference images belong to the same data distribution. The model was trained on 90% of the images for 25 epochs, and evaluated with the 10% held out set. I kept the best-performing model (at epoch 15). This is the main model used for conducting inference on input satellite images in the GiveDirectly study area.

Throughout the training process, I conducted extensive data augmentation to increase the transferability of the model from one dataset to another. I randomly flipped the training images horizontally and vertically, randomly jittered the brightness, contrast, saturation, and hue of the images. For the Open AI Tanzania dataset, I also randomly blurred and cropped the images.

C Validation in Mexico

I conducted an additional validation exercise against the 2010 Population and Housing Census in Mexico [4]. In particular, I utilized the locality-level data set, Principales Resultados por Localidad, or ITER. (A locality is equivalent to a village in rural areas.) To form the analysis sample, I dropped all urban localities (defined as having more than 2,500 residents), small localities where the relevant asset measures were masked in the census to protect privacy, and localities where these measures were missing. To avoid covering neighboring urban or rural localities in the satellite images, I excluded rural localities that were closer than 0.01 degree (1.1 km) from other rural localities, or 0.1 degree (11.1 km) from urban localities. Finally, to reduce computation, I randomly sampled 200 rural localities, and dropped 3 of them, for which Google Static Maps did not have satellite image coverage for.

I extracted three variables from the census records: asset score, durable asset score and population count. I computed an asset index to measure wealth, in line with the prior literature [5]. The asset index is 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), housing (cement floor, house with ≥ 2 bedrooms, house with ≥ 3 rooms), and public good (toilet, electricity, piped water, and drainage). An asset index was calculated for each category, and an overall asset index was calculated by pooling all the categories. When necessary, I flipped the signs of the asset indices, such that a higher score indicates more wealth.

In the census, each rural locality was geo-coded as a point. Most of the rural localities were small, isolated and surrounded by vegetation or open space, making it feasible to match census records to corresponding satellite images. For each locality, I obtained satellite images that cover an area of roughly 1×1 km, with the locality coordinate at the center. The images were retrieved from the Google Static Maps API on October 10, 2019, and were likely taken several years after the census. I generated deep learning predictions on these images with the method described in Methods and Supplementary Materials B, but only trained the model for the first three steps in Supplementary Materials B.2. From the deep learning predictions, I computed two metrics of interest: the total number of houses in a locality, and the average size of houses (winsorized at the 99 percentile). Additionally, I downloaded night light data, the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) composite images from 2019, to conduct the comparison.

References

1. He, K., Gkioxari, G., Dollár, P. & Girshick, R. *Mask R-CNN* in *Proceedings of the IEEE international conference on computer vision* (2017), 2961–2969.
2. *COCO - Common Objects in Contexts* http://cocodataset.org. accessed 6 May 2020.
3. *2018 Open AI Tanzania Building Footprint Segmentation Challenge* https://competitions.codalab.org/competitions/20100. accessed 6 May 2020.
4. *2010 Population and Housing Census of Mexico* https://www.inegi.org.mx/programas/ccpv/2010/default.html. accessed 5 May 2020.
5. Jean, N. *et al.* Combining satellite imagery and machine learning to predict poverty. *Science* **353,** 790–794 (2016).