NBER WORKING PAPER SERIES

USING SATELLITE IMAGERY AND DEEP LEARNING TO EVALUATE THE IMPACT OF ANTI-POVERTY PROGRAMS

Luna Yue Huang
Solomon M. Hsiang
Marco Gonzalez-Navarro

Working Paper 29105
http://www.nber.org/papers/w29105

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2021

We thank Edward Miguel, Jeremy Magruder, Ben Faber, Marshall Burke, Joshua Blumenstock, Supreet Kaur, Ethan Ligon, Elisabeth Sadoulet, Alain De Janvry, Aprajit Mahajan, the participants in the AGU Fall Meeting 2019 (Session GC34C), the UC Berkeley Trade Lunch, Development Workshop, Development Lunch, Good Data Seminar, and the FAO Technical Network on Poverty Analysis for feedback. We thank Edward Miguel, Michael Walker, Dennis Egger, Johannes Haushofer, Paul Niehaus, and the rest of the GiveDirectly team, for generously sharing the dataset with us and responding to our inquiries. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Using Satellite Imagery and Deep Learning to Evaluate the Impact of Anti-Poverty Programs
Luna Yue Huang, Solomon M. Hsiang, and Marco Gonzalez-Navarro
NBER Working Paper No. 29105
July 2021
JEL No. C8,H0,O1,O22,Q0,R0

ABSTRACT

The rigorous evaluation of anti-poverty programs is key to the fight against global poverty. Traditional approaches rely heavily on repeated in-person field surveys to measure program effects. However, this is costly, time-consuming, and often logistically challenging. Here we provide the first evidence that we can conduct such program evaluations based solely on high-resolution satellite imagery and deep learning methods. Our application estimates changes in household welfare in a recent anti-poverty program in rural Kenya. Leveraging a large literature documenting a reliable relationship between housing quality and household wealth, we infer changes in household wealth based on satellite-derived changes in housing quality and obtain consistent results with the traditional field-survey based approach. Our approach generates inexpensive and timely insights on program effectiveness in international development programs.

Luna Yue Huang
Agricultural and Resource Economics
University of California, Berkeley
207 Giannini Hall
Berkeley, CA 94720-3310
yue_huang@berkeley.edu

Solomon M. Hsiang
Goldman School of Public Policy
University of California, Berkeley
2607 Hearst Avenue
Berkeley, CA 94720-7320
and NBER
shsiang@berkeley.edu

Marco Gonzalez-Navarro
Agricultural and Resource Economics
UC-Berkeley
207 Giannini Hall
Berkeley, CA 94720-3310
and University of California, Berkeley
marcog@berkeley.edu

Introduction

Rigorous impact evaluation forms the basis of the modern approach to fight global poverty and provides input for evidence-based policy making (*1,2*). The impacts of anti-poverty interventions are almost universally evaluated using household surveys, typically comprehensive questionnaires containing hundreds of questions that can touch every aspect of people's lives. However, such field surveys are often prohibitively expensive to conduct (*3,4*) and unanticipated events, such as political unrest or public health crises, frequently disrupt them (*5*). In this paper, we provide the first demonstration that the household welfare impacts of a large scale anti-poverty randomized controlled trial (RCT) can be accurately measured relying solely on satellite data, instead of household surveys.

Recent advances enable poverty to be identified remotely (*6–13*) and widespread adoption of mobile phones allows targeted anti-poverty interventions to be deployed over-the-network (*14*), such as the cash transfer program we study here (*15*). By demonstrating that the impacts of such interventions can be evaluated remotely, we hope that future programs can, in principle, be designed, deployed, and evaluated with limited reliance on logistically complex and expensive ground operations. Because costs and logistics play a major role in limiting the scale of anti-poverty programs (*16*), simplifying their deployment and evaluation is crucial to achieving their full global potential.

We study a large pre-existing trial (*15*) that was recently completed and evaluated with field surveys and show that we can consistently recover the impact of the program using satellite imagery and deep learning. While previous studies have successfully evaluated the environmental impacts of randomized controlled trials with remote sensing data (*17, 18*), we are not aware of studies that demonstrate similar successes for household economic well-being. Specifically, we combine high-resolution daytime imagery (*19*) and state-of-the-art deep learning models (*20*) to measure housing quality among treatment and control households, and estimate the program effects on housing quality. We then map housing quality to household wealth for these households by inverting an "Engel curve," an established concept in economics (*21–24*) that describes household spending on specific goods as a function of economic well-being. Using this approach, we accurately recover the program effects on household wealth for a fraction of the cost ($0.006 per household, see Supplementary Text A) that would typically be spent on household surveys ($18–300 per household (*4*)).

Early work has shown that satellite data can be used to monitor economic development by correlating nighttime luminosity, i.e., the amount of light emitted from Earth at night (hereafter "night light") with Gross Domestic Product (GDP) at national and subnational scales (*25–27*). However, the night light data show poor sensitivity in less developed and rural areas (*6*), 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 (*25*). This makes the data less useful for studying the very target of many international development programs—people living under the poverty line. Additionally, the low spatial granularity of night light prevents it from being used to evaluate programs reliant on fine spatial variations, including most randomized controlled trials in which households in close proximity to one another are assigned to different treatments.

We propose an alternative approach—we analyze daytime imagery using a deep-learning model (*20*) to explicitly measure the quality of housing, a tangible and verifiable asset that is known to be a powerful proxy for household wealth. Even in communities where electrification rates are low, housing quality remains a strong predictor of wealth as it is often a family's single most valuable asset (*28*) and accounts for a sizable portion (10–20%) of total household expenditure globally (*29*). Furthermore, in most rural and low-income contexts, individuals do not trade up their residence but rather upgrade existing housing by expanding or building new structures on their property in response to improved economic conditions, making housing footprint a meaningful proxy for welfare. In this study, we focus on building footprint because it can be precisely measured at scale with modern deep learning techniques.

Many features of buildings other than footprint are observable with satellite imagery; for example roof material (*30, 31*). One of the main advantages of the method proposed here compared to alternative "black-box" machine learning approaches to measuring wealth that utilize *all* available information contained in satellite images (such as convolutional neural networks (*6, 11*) or random kitchen sinks (*32*)) is that it allows the exclusion of subsets of satellite-derived outcomes that may have been directly impacted by the intervention. We show the benefits of this feature of our method in the context of the experiment we evaluate. Specifically, households were eligible for the GiveDirectly study as long as their roofing was of low quality (thatched). Due to this eligibility criterion, treatment households were "prompted" to use the GiveDirectly transfer to upgrade their roofing as a way to signal to the experimenters that they had used the cash for good. An improvement of roofs among participating households beyond what would have been expected solely from wealth increases biases estimates of wealth when methods cannot exclude subsets of outcomes. In contrast, it is straightforward for our method to focus exclusively on subsets of available information that were not affected directly (in this case building footprints) while ignoring problematic outcomes (such as roof material) in order to provide unbiased estimates of wealth effects.

Results

We evaluate a development intervention that was conducted in 2014–2017 in 653 villages in rural Kenya (*15*). GiveDirectly, a US charity, implemented a randomized controlled trial of unconditional cash transfers to rural households via mobile money, using as sole eligibility criterion whether the household lived under a thatched roof (a low quality roof material that served as a simple means test). Each treatment household received $1,000—equivalent to about 75% of annual household expenditures—in lump sum, and could spend it however it wished. To evaluate the effectiveness of the program, GiveDirectly randomly selected 328 villages as the treatment group, where eligible households (about 1/3 of the population) received transfers, and used the remaining 325 villages as the control group. The authors conducted extensive household surveys before and after the distribution of the transfers to measure program impacts as is the current practice in the evaluation literature.

Mapping Treatment Intensity and Housing Quality. To evaluate program impacts, we first construct a map that shows the intensity of the anti-poverty program (hereafter "treatment") in different geographical units (in this case it is simplest to work with raster grid cells). This geocoded information is obtained from program implementation records, which document where the program was administered. Because of the extremely high granularity of satellite-derived housing quality metrics, it is feasible to study programs that induce fine spatial variation such as household-level 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 program evaluation project.

For the GiveDirectly experiment, we construct the treatment intensity map from a local census fielded in 2014–2015, which surveyed all the 65,385 households living in the study area (*15*). The census data record each household's geo-location, and indicate 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. Our sample contains 11,055 treatment households and 10,682 control households in total. We lay out a regular grid, and count 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 our preferred measure of treatment intensity (Figure 1c).

Next, we measure housing quality in daytime satellite images with deep learning techniques. The input images are from Google Static Maps (*19*). They are taken after the GiveDirectly intervention, have a spatial resolution of about 30cm per pixel, and contain only the RGB (red, green, blue) bands (Figure 1d).

To segment buildings, we train a state-of-the-art deep learning model, Mask R-CNN (*20*), on large, publicly available datasets such as COCO (Common Objects in Context) (*33*) and Open AI Tanzania (*34*), as well as a small annotated dataset, which are randomly sampled from all the input images (see Supplementary Text B for details on model training). The model predictions are highly accurate, both quantitatively (Figure S1) and qualitatively (Figure S2). The model generalizes well to other countries, such as Mexico, where the number of houses identified in the deep learning predictions is highly correlated with the census population count (Figure S9 and Supplementary Text C). After post-processing, each predicted instance of buildings is represented by a polygon and a "representative" roof color (Figure 1e). 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, we can measure housing outcomes for each household.

We extract two metrics for each built structure: the size of building footprint, and the type of roof material. The roofs are classified into three types: tin roof, thatched roof, and painted roof, based on their color profiles (Figure S3). Compared to tin roofs, thatched roofs are generally of lower quality (*15, 35*). (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 (*30, 31*). As such, we aggregate the total building footprint to measure all housing assets (Figure 1f, Building Footprint), and the footprint of tin-roof buildings to measure high-quality housing assets (Figure 1f, Tin-roof Area), in each grid cell. To obtain night light data for systematic comparison, we download and resample the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) composite images in 2019 (*36, 37*).

The maps of treatment intensity and remotely sensed outcomes for the GiveDirectly experiment are shown in Figure 2. For visual display and privacy protection purposes, we 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). Both of the housing quality measures capture richer variation in the entire area (Figure 2b, c), whereas the night light data demonstrate little variation in this rural, sparsely populated area, except in a few spots close to local towns (Figure 2d).

Estimating the Program Effects on Housing Quality. We regress the remotely sensed outcomes on treatment intensity to estimate the causal effects of the GiveDirectly cash transfer. We choose a spatial resolution of 0.001° (approximately 100m), such that most of the grid cells contain 0–5 households. We exploit only the experimentally-induced random variation in treatment intensity for identification, and account for pre-determined differences in program eligibility. Intuitively, consider two grid cells, one containing a household that received the transfer, and the other containing a household that was eligible to get the transfer but did not because it was randomized into the control group. With valid randomization (*15*), the differences in outcomes between the two can be attributed to the cash transfer. We plot the causal effects on night light and housing quality as cash infusion intensity increases (Figure 3, in color), without making assumptions on the structure of the effects. The results suggest that the effects grow linearly with the amount of cash infusion. We therefore also report an "average" effect, estimated with the assumption that each $1,000 transfer generates an effect of the same magnitude (Figure 3, panel subtitles). We demonstrate the validity of the empirical strategy further by running 100 placebo simulations—we artificially generate placebo cash transfers that did not actually take place but is consistent with the original randomization design, and estimate their treatment effects (Figure 3, in gray). The resulting estimates are reassuringly centered around zero.

We observe statistically significant and economically sizable effects on housing quality, on both the extensive margin (larger building footprint) (Figure 3a), and the intensive margin (higher quality roofs) (Figure 3b). On average, a $1,000 cash transfer significantly increased building footprint by 7.9 square meters (95% CI: [2.3, 13.5], $t(14, 143) = 2.8$, $p = 0.006$) or 85.0 square feet, and tin-roof area by 13.6 square meters (95% CI: [9.6, 17.6], $t(14, 143) = 6.7$, $p < 0.001$) or 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 (*35*). These estimates are consistent with the results from extensive field surveys, which also documented large increases in housing asset values (*15*).

On the other hand, we do not observe any program effects on night light (Figure 3c), despite the fact that the cash transfer had large positive impacts on many aspects of the recipient households' economic well-being—food expenditure, consumer durable spending, asset holding, and housing values (*15*). The estimated effect is -0.000120 (95% CI: $[-0.008, 0.008]$, $t(14, 143) = -0.03$, $p = 0.977$) which is not statistically different from zero, is small in magnitude, and actually slightly negative. This may be because of low demand for electrification (*38*), or the poor sensitivity of night light in low-income, rural regions (*6*).

Recovering the Program Effects on Economic Well-being with Engel Curves. We recover the program effects on household economic well-being with a canonical economic concept, the Engel curve. Engel curves describe how household expenditures on particular goods or services are related to households' economic well-being. For example, it is widely known that poorer families spend a larger share of their expenditure on food. Engel curves have long been used to infer economic well-being without needing detailed information on prices as it is straightforward to measure how much of a household's expenditure is spent on food (*21–24*). We adapt this concept to housing quality by exploiting the fact that someone who lives in a larger house is likely to be wealthier than someone who lives in a smaller house (Figure 4a). By the same logic, if we observe that someone's house size increased, then we can infer what level of wealth is associated to such a house size—as if they were moving up on the Engel curve. Mathematically, the slope of the Engel curve represents the ratio between the change in house size and the change in wealth. We divide the change in the house size (Figure 3) by the slope of the Engel curve (Figure 4a) to infer the corresponding change in wealth (Figure 4b). Importantly, the validity of this approach depends on the assumption that the Engel curve does not shift in response to the treatment, which could happen due to relative price changes of the good or taste changes.

In this study, we derive housing Engel curves from an endline survey of the GiveDirectly trial participants between May 2016 and June 2017, which includes 4,578 geo-coded households who were eligible for the transfer. Of these households, only those assigned to the control group are used for the estimation. In Figure 4a, we show the relationship between survey-based measures of economic well-being (x-axis) and remotely sensed night light or housing quality measures (y-axis). The Engel curves are estimated with a linear regression (dotted lines). The non-linear fit with LOESS (solid lines) shows only small deviations from the linear regression line, and we cannot reject the null hypothesis that these Engel curves are linear (see Materials and Methods). The Engel curves are also roughly monotonically increasing, validating the choice of these variables as wealth proxies.

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 (LSMS)) to estimate Engel curves. We demonstrate this by comparing the Engel curves derived from two distinct samples: the households who were deemed eligible to receive the cash transfers (meaning that they used to live in thatched-roof houses), and households who were not. While all the households live in the same area in western Kenya, the ineligible households are generally wealthier than the eligible ones. Their Engel curves, however, are similar within the same range of wealth (Figure S8).

We scale program effects on each remotely sensed outcome by the Engel curve slope to estimate the impacts of the GiveDirectly transfer on household wealth, measured by aggregating the values of a variety of assets as measured with household surveys. In Figure 4b, we compare the satellite-derived estimates against the survey-based estimates, which are computed from rich endline household survey data and taken from Table 1, Column 1 in the original paper (*15*). As can be seen, the estimate based on building footprint (USD 425 PPP, 95% CI: [61, 788]) is informative and very close to the survey based estimate (USD 556 PPP, 95% CI: [485, 626]). For reference, the entire GiveDirectly cash transfer is worth USD 1,871 PPP (USD 1,000 nominal). Note that the estimate based on night light is slightly negative and imprecise, and both the upper and lower bounds are uninformative. In contrast, the estimate based on tin-roof area is about two times as large as the survey-based estimate. The results are qualitatively similar when we distinguish between housing asset (Figure S4) and non-housing asset (Figure S5), or when we use annual consumption expenditure as the alternative measure of economic well-being (Figure S6).

Why is the estimate based on the tin-roof area much larger than the survey based estimate? We argue this is due to the violation of a key assumption, which is that the Engel curve used to estimate changes in wealth cannot change directly in response to the treatment—only through its wealth effects. To give intuition for why this matters, consider a program that directly gives people food. In such a case we can no longer look at food consumption to infer program effects on economic well-being, because the relationship between the food and income will be altered directly by the program and households will "look" wealthier than they really are based on their food consumption. More relevant for impact evaluation using satellite data, this example is analogous to examining the impacts of a program that provides roads to a region. One would need to exclude the program roads themselves contained in satellite images and look at other correlates of welfare to estimate impacts of such a roads program in an unbiased manner. In the GiveDirectly case, only households that lived in thatched-roof houses were eligible for the study. Households' usual consumption patterns of high-quality tin roofs might have been affected by this eligibility criteria. One can observe that treatment households owned more tin-roof buildings compared to control households with the same amount of wealth (Figure S7). This may have been a result of households interpreting the treatment as a "labelled" cash transfer (*39*).

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 (*6–13*). 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 economic well-being from observing the consumption patterns of these private or public goods (similar 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 in this study), introducing biases in the estimated program effects. In this study, we disentangle the two tasks, so that the first task can be framed as a traditional object detection and segmentation task, allowing us to leverage extensive research in computer science; and the second task becomes more transparent, explicit, and the assumptions testable (for example, with Figure S7).

Discussion

This paper provides evidence that RCT program evaluations aimed at improving household welfare can be obtained solely based on satellite imagery and deep learning methods. This approach has the advantage of being inexpensive and timely, suggesting great promise as a complement and in some cases as a substitute to in-person survey data collection methods.

However, it bears noting that a fundamental limitation to evaluating programs based on satellite imagery is that in order to be measurable from space, programs being evaluated have to generate impacts on the built landscape. This prevents applicability to programs targeted at addressing development challenges that are unlikely to impact the built environment such as improved teaching methods at schools. Another limitation is that welfare is a household or individual concept whereas satellite images capture characteristics about a place. Mapping household welfare to housing as we do here requires a tight mapping between structures and households through limited mobility. While migration rates are very low in the GiveDirectly study area (*15*), this may be a challenge for programs that impact mobility, such as transportation infrastructure programs.

References

1. A. Deaton, *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy* (The World Bank, 1997).

2. A. V. Banerjee, E. Duflo, *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty* (Public Affairs, 2011).

3. S. Pamies-Sumner, Development impact evaluations: State of play and new challenges, *Tech. rep.*, Agence Française de Développement (2015).

4. J. M. Alix-Garcia, K. R. Sims, L. Costica, Better to be indirect? Testing the accuracy and cost-savings of indirect surveys, *World Development* **142**, 105419 (2021).

5. 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).

6. N. Jean, *et al.*, Combining satellite imagery and machine learning to predict poverty, *Science* **353**, 790-794 (2016).

7. J. E. Blumenstock, Fighting poverty with data, *Science* **353**, 753-754 (2016).

8. R. Engstrom, J. Hersh, D. Newhouse, Poverty from space: Using high-resolution satellite imagery for estimating economic well-being, *Tech. rep.*, The World Bank (2017).

9. B. Babenko, J. Hersh, D. Newhouse, A. Ramakrishnan, T. Swartz, *Proceedings of NIPS 2017 Workshop on Machine Learning for the Developing World* (2017).

10. G. R. Watmough, *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).

11. 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).

12. E. L. Aiken, G. Bedoya, A. Coville, J. E. Blumenstock, Targeting development aid with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan (2020). Unpublished.

13. J. Blumenstock, Machine learning can help get covid-19 aid to those who need it most, *Nature* (2020).

14. T. Suri, Mobile money, *Annual Review of Economics* **9**, 497-520 (2017).

15. 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).

16. C. Blattman, P. Niehaus, Show them the money: Why giving cash helps alleviate poverty, *Foreign Affairs* **93**, 117-126 (2014).

17. J. Alix-Garcia, C. McIntosh, K. R. E. Sims, J. R. Welch, The ecological footprint of poverty alleviation: Evidence from Mexico's oportunidades program, *Review of Economics and Statistics* **95**, 417-435 (2013).

18. S. Jayachandran, *et al.*, Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation, *Science* **357**, 267-273 (2017).

19. Google Static Maps, Google static maps, https://developers.google.com/maps/documentation/maps-static/intro (2020). Accessed 6 May 2020.

20. K. He, G. Gkioxari, P. Dollár, R. Girshick, *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2961–2969.

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

22. A. Tarozzi, A. Deaton, Using census and survey data to estimate poverty and inequality for small areas, *Review of Economics and Statistics* **91**, 773-792 (2009).

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

24. D. Atkin, B. Faber, T. Fally, M. Gonzalez-Navarro, Measuring welfare and inequality with incomplete price information, *Tech. rep.*, National Bureau of Economic Research (2020).

25. J. V. Henderson, A. Storeygard, D. N. Weil, Measuring economic growth from outer space, *American Economic Review* **102**, 994-1028 (2012).

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

27. S. Michalopoulos, E. Papaioannou, National institutions and subnational development in Africa, *Quarterly Journal of Economics* **129**, 151-213 (2014).

28. M. Gonzalez-Navarro, C. Quintana-Domeque, The reliability of self-reported home values in a developing country context, *Journal of Housing Economics* **18**, 311-324 (2009).

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

30. 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).

31. G. Michaels, *et al.*, Planning ahead for better neighborhoods: Long run evidence from tanzania, *Journal of Political Economy* **129**, 2112–2156 (2021).

32. E. Rolf, *et al.*, A generalizable and accessible approach to machine learning with global satellite imagery, *Tech. rep.*, National Bureau of Economic Research (2020).

33. COCO, Coco - common objects in contexts, http://cocodataset.org (2020). Accessed 6 May 2020.

34. Open AI Tanzania, 2018 open AI Tanzania building footprint segmentation challenge, https://competitions.codalab.org/competitions/20100 (2020). Accessed 6 May 2020.

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

36. Google Earth Engine, 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 (2020). Accessed 6 May 2020.

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

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

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

40. D. Lindenbaum, 2nd spacenet competition winners code release, https://medium.com/the-downlinq/2nd-spacenet-competition-winners-code-release-c7473eea7c11 (2017). Accessed 29 Jan 2021.

41. S. Mills, S. Weiss, C. Liang, *Earth Observing Systems XVIII* (International Society for Optics and Photonics, 2013), vol. 8866, p. 88661P.

42. 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).

43. T. G. Conley, Gmm estimation with cross sectional dependence, *Journal of Econometrics* **92**, 1-45 (1999).

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

45. 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).

46. 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 (2016). Accessed 10 May 2020.

47. Google Static Maps, Maps static API usage and billing, https://developers.google.com/maps/documentation/maps-static/usage-and-billing (2021). Accessed 20 April 2021.

48. INEGI, 2010 population and housing census of Mexico, https://www.inegi.org.mx/programas/ccpv/2010/default.html (2010). Accessed 5 May 2020.

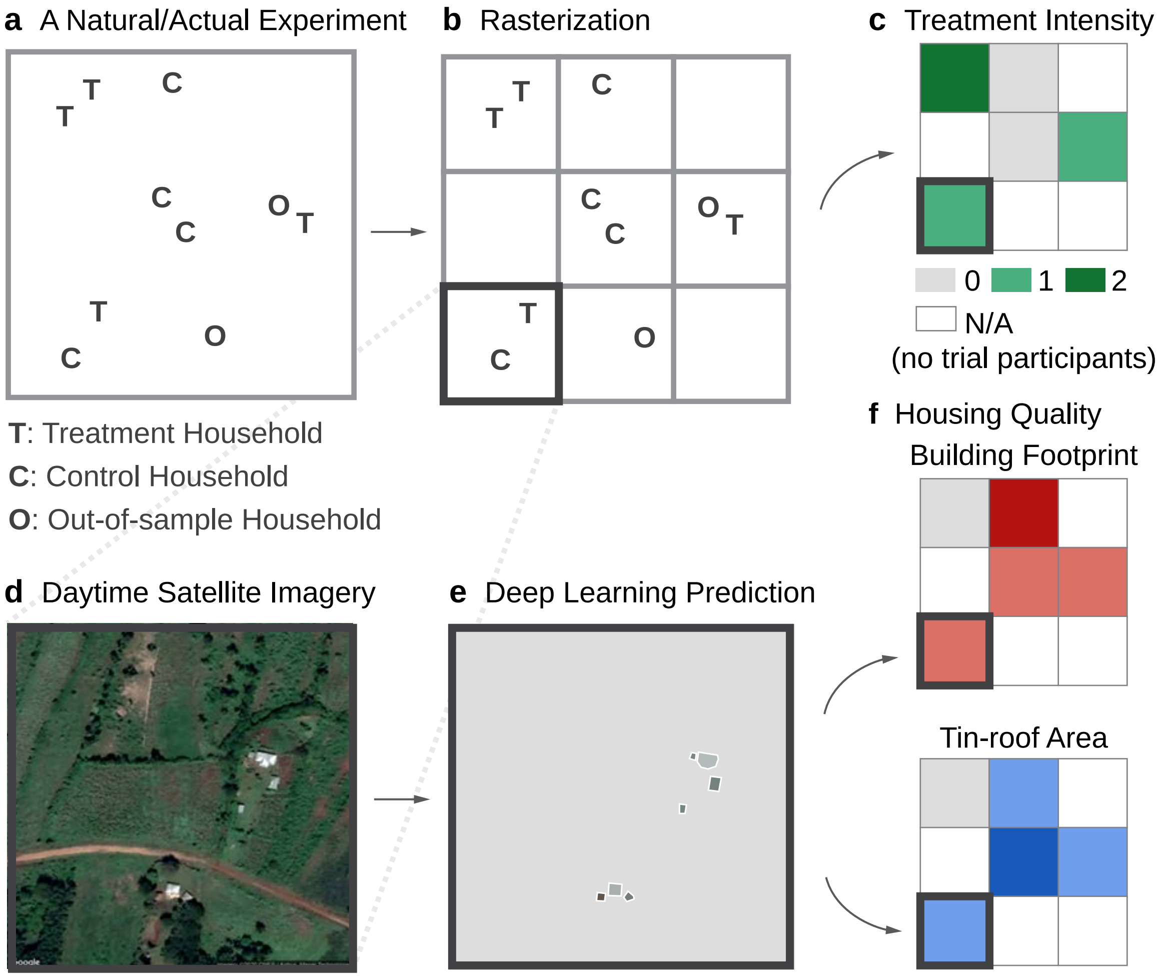

Figure 1: **Constructing maps of treatment intensity and remotely sensed outcomes from program implementation records and satellite imagery.** **a** An illustration of geocoded program implementation records. **b** Placing a regular grid over **a** and measuring the intensity of the treatment in each grid cell. **c** Constructed raster of the number of treatment households in each grid cell. **d** An example daytime satellite image from Google Static Maps. **e** Example deep learning predictions on **d**. Each building is outlined in white and filled with the "representative" roof color. **f** Constructed rasters of remotely sensed housing quality outcomes. In **c** and **f**, grid cells without trial participants are omitted and shown in white.

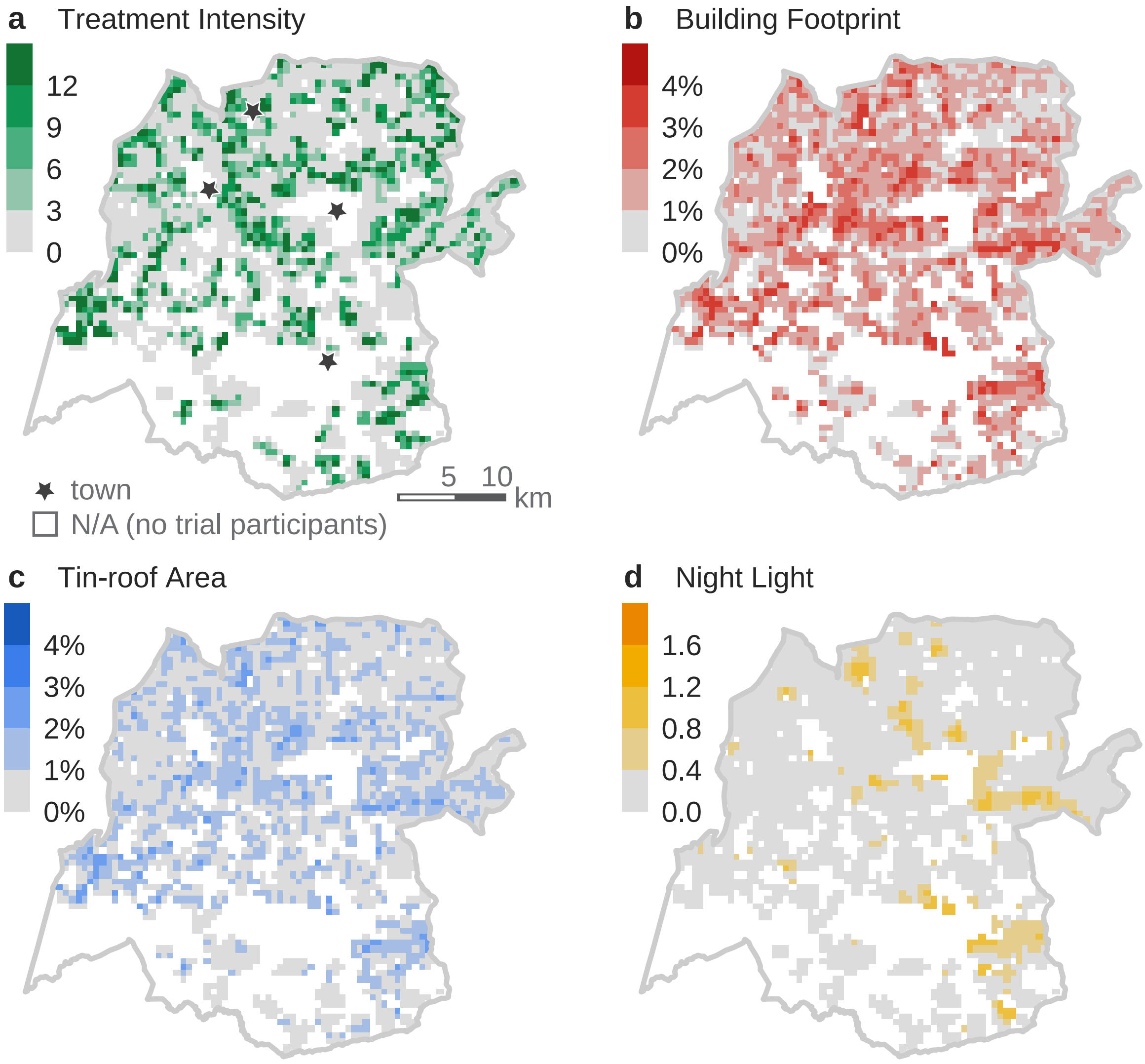

Figure 2: **Mapping treatment intensity and remotely sensed outcomes in the GiveDirectly study area in 2019. a** Treatment intensity represents the number of households who received a $1,000 cash transfer from GiveDirectly. **b** Building footprint measures the total area covered by any building, shown as a percentage of the total area. **c** Tin-roof area measures the total footprint of buildings with roofs made of tin (a high quality construction material), shown as a percentage of the total area. **d** Night light is the average radiance in the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB). In all the panels, the gray lines outline the GiveDirectly study area in Siaya, Kenya. Grid cells without trial participants are omitted and shown in white. $n = 2,501$.

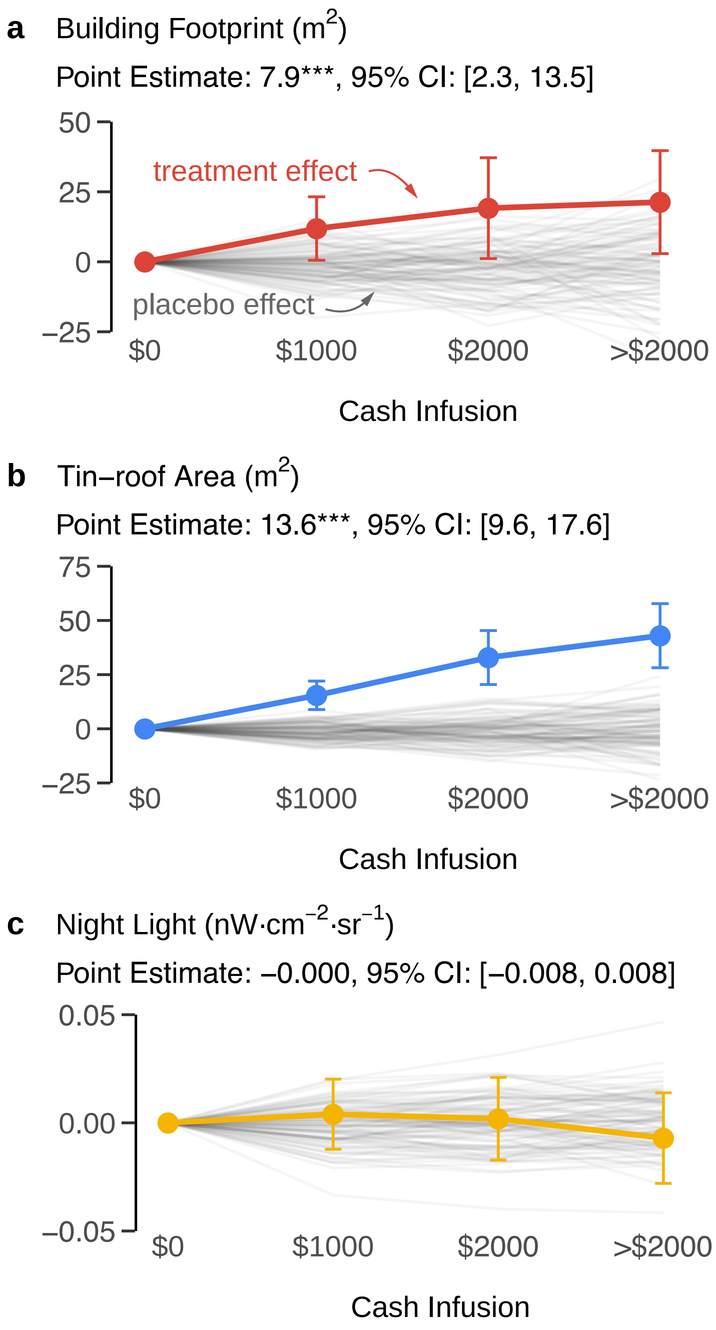

Figure 3: **Housing quality increased in response to the GiveDirectly cash transfer, but night light remained unchanged.** The treatment effects of the cash transfers on building footprint (**a**), tin-roof area (**b**), and night light (**c**) are shown in color. The dots represent the point estimates, and the error bars represent the 95% confidence intervals. Gray lines show the estimated effects of the placebo cash infusions from 100 simulations. The panel subtitles report the average treatment effect of a $1,000 transfer and the 95% confidence intervals, assuming constant effect. *** indicates statistical significance at the 1% level for a two-sided t-test. $n = 14,155$.

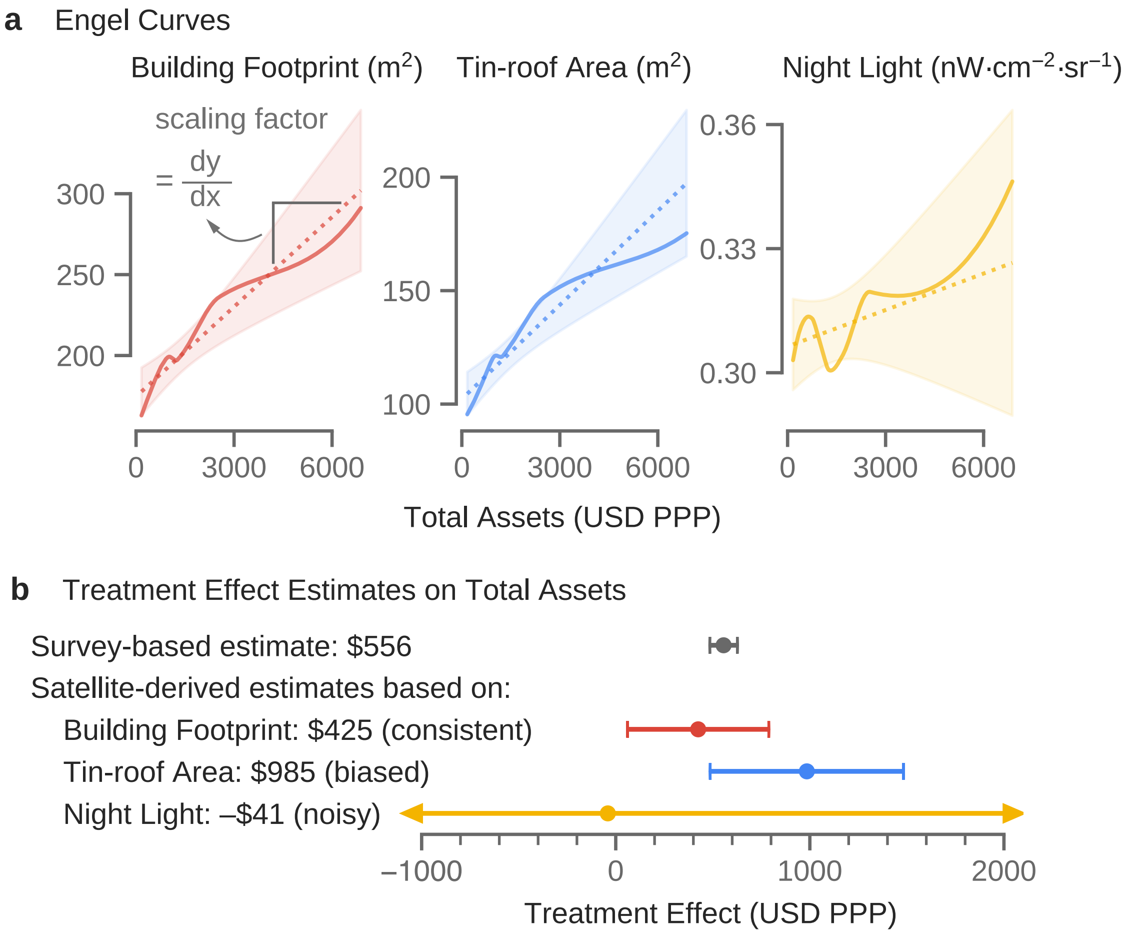

Figure 4: **The treatment effect on total assets can be correctly recovered by scaling the effect on building footprint. a** The Engel curves of building footprint, tin-roof area, and night light, estimated with LOESS (solid line) or a linear regression (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the survey-based versus satellite-derived treatment effects. The dots show the point estimates. The error bars show the 95% confidence intervals, with the arrow(s) marking upper/lower bounds that are out of range (if any). $n = 1,844$.

Supplementary Materials for

Using Satellite Imagery and Deep Learning to Evaluate the Impact of Anti-Poverty Programs

Luna Yue Huang,[1,2,*] Solomon Hsiang,[2,3] Marco Gonzalez-Navarro[1]

[1]Department of Agricultural and Resource Economics, UC Berkeley, Berkeley, CA, USA
[2]Global Policy Laboratory, Goldman School of Public Policy, UC Berkeley, Berkeley, CA, USA
[3]National Bureau of Economic Research, Cambridge, MA, USA

[*]Corresponding author. E-mail: yue_huang@berkeley.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figure S1 to S9

Materials and Methods

Constructing the Treatment Intensity Map. To construct the treatment intensity map, we utilize 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 (*15*). To address the measurement errors of the GPS collection devices, we discard 58 outliers (living more than 2 kilometers away from the village centers) and impute those and other 4 missing GPS coordinates with village center coordinates. Then, we convert these household records into a raster map. We lay out a regular grid, and count, 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 are excluded. To account for pre-determined policy intensity differences, we record (and later control 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. We utilize high-resolution daytime satellite images from Google Static Maps (*19*). These images have a spatial resolution of about 30cm per pixel (at equator), and contain only the RGB (red, green, blue) bands (see Figure 1d and 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 we 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. We first leverage a state-of-the-art deep learning model, Mask R-CNN (*20*) to segment buildings—that is, to detect each building and the pixels that they occupy—in the Google Static Map satellite images.

We then convert the pixel-wise predictions to polygons, and extract housing quality metrics related to the size of the building and the roof materials from each polygon (see Figure 1e and 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.

We train the Mask R-CNN model with a multi-step process and a transfer learning framework, as described in greater detail in Supplementary Text 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. We curate 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. We pre-train the Mask R-CNN model on large, publicly available datasets such as COCO (Common Objects in Context) and Open AI Tanzania, and fine-tune 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 (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 (*40*). 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.

We post-process 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, we compute two housing quality metrics: building footprint and type of roof materials. We then lay out a regular grid, assign each building to grid cells based on the centroids of the polygons, and aggregate to obtain two metrics at the pixel level: building footprint (Figure 2b) and tin-roof area (Figure 2c).

First, we measure the size of each building polygon and convert it to square meters. We correct 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, we estimate the types of roof materials based on the colors of the roofs, and compute the footprint of tin-roof buildings in each grid cell. For each building, we take all the pixels associated with the given building instance, and assign 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, we project all the RGB color vectors to the CIELAB color space, and cluster these roof color vectors into 8 groups by running the K-means clustering algorithm. We further classify these 8 groups into three types of roof materials: tin roof, thatched roof, and painted roof (Figure S3), and compute the total footprint of tin-roof buildings.

Obtaining the Night Light Data. To measure nighttime luminosity, we use the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) composite images hosted on Google Earth Engine (*36, 37*). The VIIRS-DNB data product excludes areas impacted by cloud cover and correct for stray light (*41*). 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 (*36*). This data product has a native spatial resolution of 15 arc seconds (approximately 463 meters at the equator), and we resample the data by conducting nearest neighbor interpolation when necessary. We average over all the monthly observations in 2019 and construct a single cross sectional observation, to reduce seasonality effects and for consistency with the daytime satellite imagery (Figure 2d). 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 (*42*). This ensures that we 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 $0.001° \times 0.001°$ grid cell (approximately 100m×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, we flexibly control for the number of eligible households per grid cell, and exclude grid cells with no eligible households. Because the grid cells are fairly small and the number of observations for $k > 2$ is small, we bin the number of recipient households into four bins $k \in K = \{0, 1, 2, 2+\}$, to preserve statistical power. Standard errors are calculated à la Conley, with a uniform kernel and a 3km cutoff (*43–46*). To reduce the effects of outliers (due to sensor malfunctioning or machine learning model prediction errors), we winsorize all remotely sensed variables at the 99 percentile.

We run 100 placebo simulations to further demonstrate the validity of the main specification. In each simulation, we randomly assign 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, we randomly assign 2/3 of the villages to the treatment group (and the rest to the control group); whereas in the low-saturation group, we assign 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 (*15*). Using these simulated placebo treatment status variables, we estimate the placebo treatment effects with the econometric specification described in Equation 1.

To compute a single pooled treatment effect, we make an assumption of linear treatment effects—every transfer of \$1,000 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. We conduct two-sided t-tests to assess statistical significance.

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' economic well-being from the consumption patterns of a limited subset of goods (*21–24*). 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 economic well-being) 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, we estimate the Engel curves—the relationships between remotely sensed metrics and survey-based measures of economic well-being—based on the endline survey of the original GiveDirectly trial, which includes a representative set of 4,578 geo-coded 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, we observe 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. We do not study land values because they are difficult to value given thin local markets (*15*).

We perform heuristic matching between the buildings and the household survey GPS coordinates, to link variables in the survey with remotely sensed variables. First, we take the baseline census data, which geo-coded every single household who lived in the study area, and assign 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, we match 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. We similarly assign 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, we keep 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 1,904 treatment households and 1,844 control households.

The Engel curves are estimated with only the control group (Figure 4a and Figure S4a, S5a and S6a). They are estimated both non-linearly with LOESS (see Equation 3 and the solid lines in Figure 4a) and linearly (see Equation 4 and the dotted lines in Figure 4a). When fitting LOESS, we allow for locally-fitted quadratic polynomials, and use 75% of the data points for each fit. We test for the non-linearity of the Engel curves in a separate procedure. We first run a linear regression, take the residuals, and fit the residuals with a natural (cubic) spline with 5 knots. We then conduct a two-sided F-test on the coefficients of the natural spline basis, and reject the null hypothesis (linearity) if these coefficients are jointly significant. We cannot reject linearity for any of the three proxies in Figure 4 (building footprint: $F(1,838) = 0.37$, $p = 0.829$; tin-roof area: $F(1,838) = 0.79$, $p = 0.533$; night light: $F(1,838) = 0.39$, $p = 0.814$). To minimize the influence of outliers, we winsorize annual expenditure, housing assets, non-housing assets and total assets at the 1 and 99 percentile of the eligible and non-eligible sample, respectively. We winsorize at the 1 percentile as outliers with a large amount of debt exist and could potentially drive the results otherwise. We similarly winsorize all the remotely sensed variables at the 99 percentile for the eligible and non-eligible sample. We exclude a small number of renters who do not own any housing assets (31 treatment households, 32 control households, and 55 ineligible households), to simplify the interpretation of the Engel curves.

Recovering the Program Effects on Economic Well-being. We adapt a prior mathematical formulation that uses the Engel curve to infer changes in economic well-being (*23*).

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 (*23*), 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 (*22*).

We estimate the treatment effects on wealth (or other measures of economic well-being) according to Equation 5 and Equation 6, with the treatment effect estimates for remotely sensed variables, and the slopes of the Engel curves. We compare the satellite-derived estimates against the survey-based estimates, taken from Table 1, Column 1 in the original paper (*15*), which were based on the endline household survey data (Figure 4b).

Supplementary Text

A Cost Estimation

We estimate that our evaluation approach costs \$0.006 per household, when accounting for imagery acquisition and computing costs. The Google Static Maps API charges users \$0.002 per image request (*47*). We estimate that our computing cost is roughly \$0.004 per household. Our entire data pipeline can be run within 72 hours on an NC24 instance with 4 K80 GPUs on Microsoft Azure, which costs \$3.60 per hour, and we have analyzed over 60,000 households. This is a liberal estimate that accounts for image downloading, model training, model inference, model validation, and regression analysis. Notably, we do not include labor costs for research and development, as these only need to be incurred once, are not relevant for application of the method, and that such labor costs are difficult to quantify.

B Training the Deep Learning Model

B.1 Creating In-sample Building Footprint Annotations

We create 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, we randomly sample 120 images for annotation. We use the Supervisely image annotation web platform to create annotations. On any given image, we 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 annotated, although they are 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

We use the Mask R-CNN model (*20*) for instance segmentation of buildings on satellite images. The backbone architecture used is ResNet50 with the Feature Pyramid Networks. The model is trained with a learning rate of 5×10^{-4} and a batch size of 10. Optimization is conducted with the Adam optimizer. We implement the deep learning pipeline with Python and PyTorch. In particular, we use the official Torchvision implementation of Mask R-CNN. We train 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 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 (*33*). Despite the fact that input images and object categories in COCO are different from target satellite images, pretraining 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 is 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 (*34*). 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 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 are a large number of unfinished structures in Zanzibar. Most input satellite images in this study contain very few unfinished structures, so we collapse the first two categories into one and drop the third category. The native resolution of the drone images is 7cm, and we down-sample the images to about 30cm to match with the resolution of the target satellite images.

In training time, 90% of the data are used for training, and the remaining 10% for validation. In order to guard against overfitting, and choose the best model, in each epoch, we evaluate 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 is trained for 50 epochs, and the best model (at epoch 43) is saved and loaded in subsequent steps.

3. Supplementary Annotations in Mexico, Tanzania and Kenya The model is 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 Figure S9. Some of these are historical images with lower data quality and more cloud coverage. These images are pooled and randomly split into a training set (90%) and a validation set (10%). The model is trained for 25 epochs, and achieves the best performance at epoch 17.

4. In-sample Annotations Finally, the model is 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 is trained on 90% of the images for 25 epochs, and evaluated with the 10% held out set. We keep 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, we conduct extensive data augmentation to increase the transferability of the model from one dataset to another. We randomly flip the training images horizontally and vertically, randomly jitter the brightness, contrast, saturation, and hue of the images. For the Open AI Tanzania dataset, we also randomly blur and crop the images.

C Validation in Mexico

C.1 Results

We provide additional validation results in rural Mexico, using the 2010 Population and Housing Census (*48*). Population count in a rural village (as reported in the 2010 census), is highly correlated with the number of houses in that village (as identified by the deep learning model), with a Pearson correlation coefficient of 0.82 (Figure S9b). Population count, however, is only modestly correlated with night light (Figure S9a). Night light is less sensitive in smaller, less populated villages, a finding that is consistent with prior work (*6*).

C.2 Methods

This comparison is based on 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, we drop all urban localities (defined as having more than 2,500 residents), small localities where the relevant asset measures are masked in the census to protect privacy, and localities where these measures are missing. To avoid covering neighboring urban or rural localities in the satellite images, we exclude rural localities that are 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, we randomly sample 200 rural localities, and drop 3 of them, for which Google Static Maps does not have satellite image coverage for.

In the census, each rural locality is geo-coded as a point. Most of the rural localities are small, isolated and surrounded by vegetation or open space, making it feasible to match census records to corresponding satellite images. For each locality, we obtain satellite images that cover an area of roughly 1×1 km, with the locality coordinate at the center. The images are retrieved from the Google Static Maps API on October 10, 2019, and are likely taken several years after the census. We generate deep learning predictions on these images with the method described in Materials and Methods and Supplementary Text B, but only train the model for the first three steps in Supplementary Text B.2. For the comparison, we count the number of houses in a locality in the deep learning predictions, and extract the population count variable from the census. Additionally, we download night light data, the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) composite images from 2019.

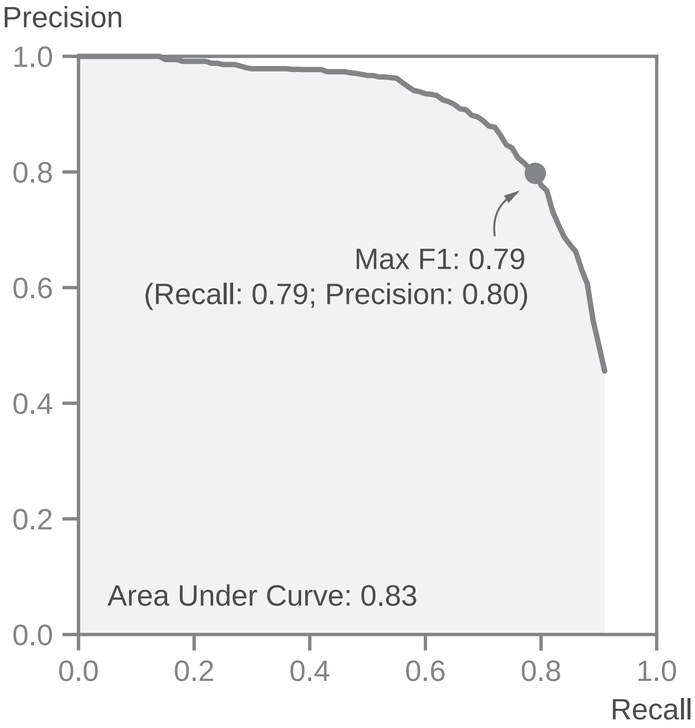

Figure S1: **The precision-recall curve of the Mask R-CNN model shows satisfactory predictive performance.** The Mask R-CNN model is trained and evaluated with 3-fold cross validation. The evaluation is based on 120 annotated images, which were randomly sampled from all the input satellite images in 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.

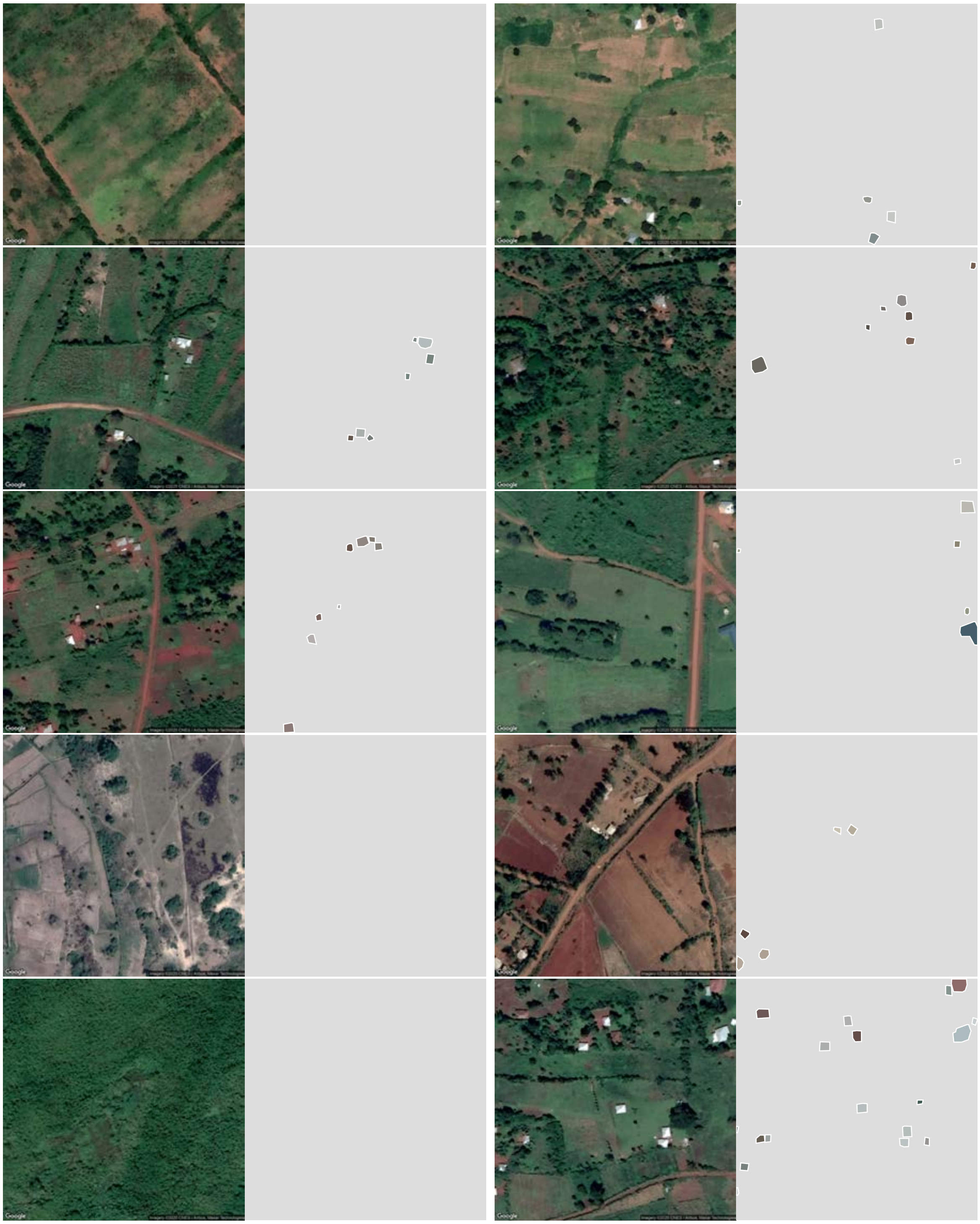

Figure S2: **Ten randomly sampled pairs of input images and deep learning predictions.** Ten images are randomly sampled from all the input satellite images in the GiveDirectly study area. Each predicted building is outlined in white and filled with the "representative" roof color.

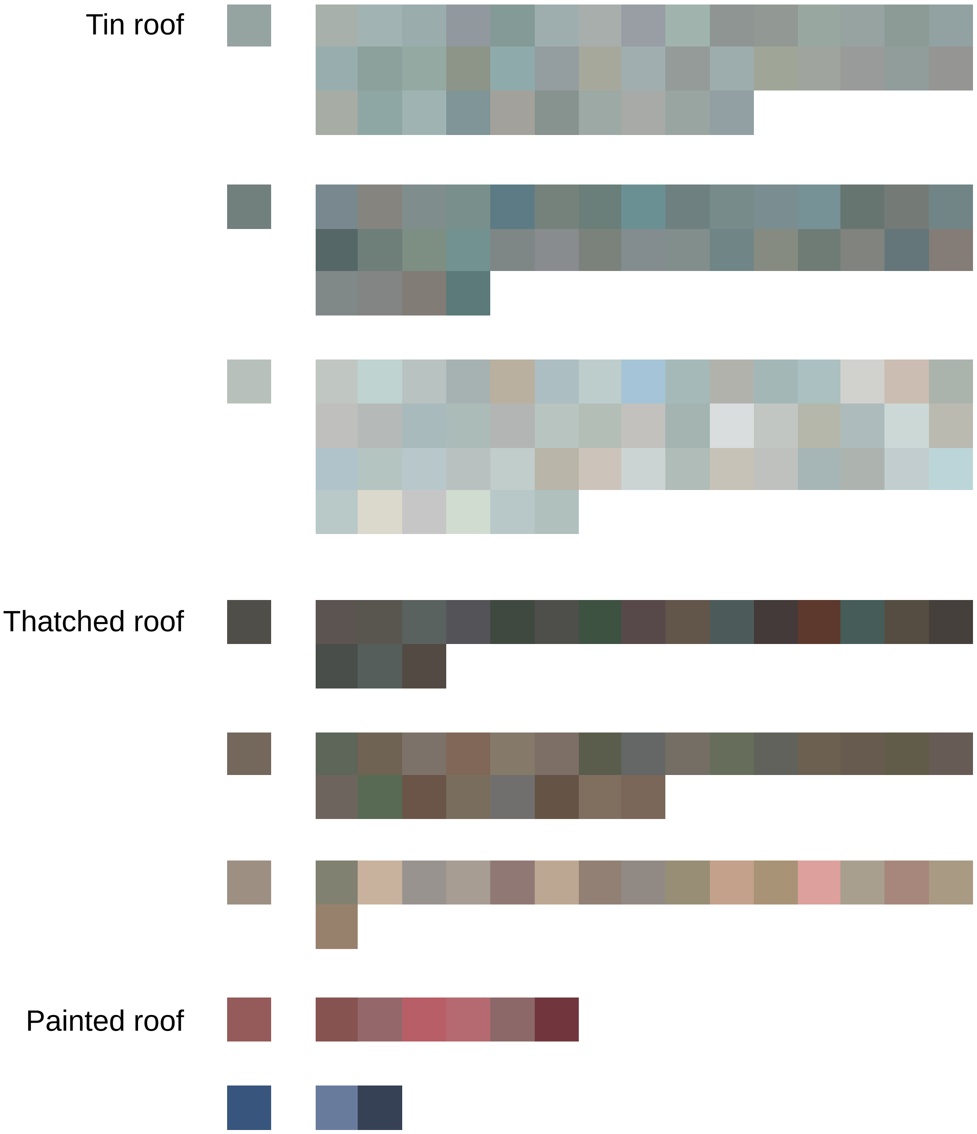

Figure S3: **The distribution and grouping of roof colors.** All the buildings in the GiveDirectly study area are split into eight groups by a K-means clustering algorithm, based on their roof colors. The color block on the left represents the “average” roof color of the cluster, and the color blocks 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.

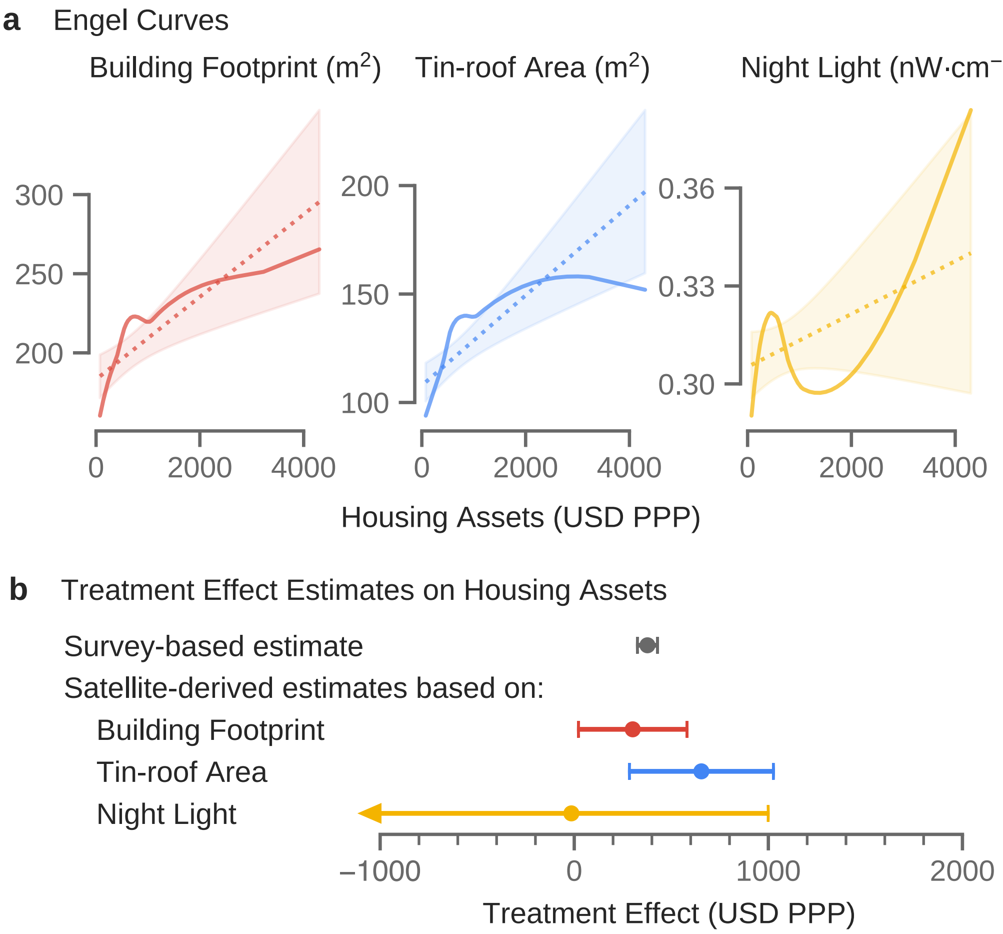

Figure S4: **The treatment effect on housing assets can be similarly recovered by scaling the effect on building footprint. a** The Engel curves of building footprint, tin-roof area, and night light, estimated with LOESS (solid line) or a linear regression (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the survey-based versus satellite-derived treatment effects. The dots show the point estimates. The error bars show the 95% confidence intervals, with the arrow(s) marking upper/lower bounds that are out of range (if any). $n = 1,844$.

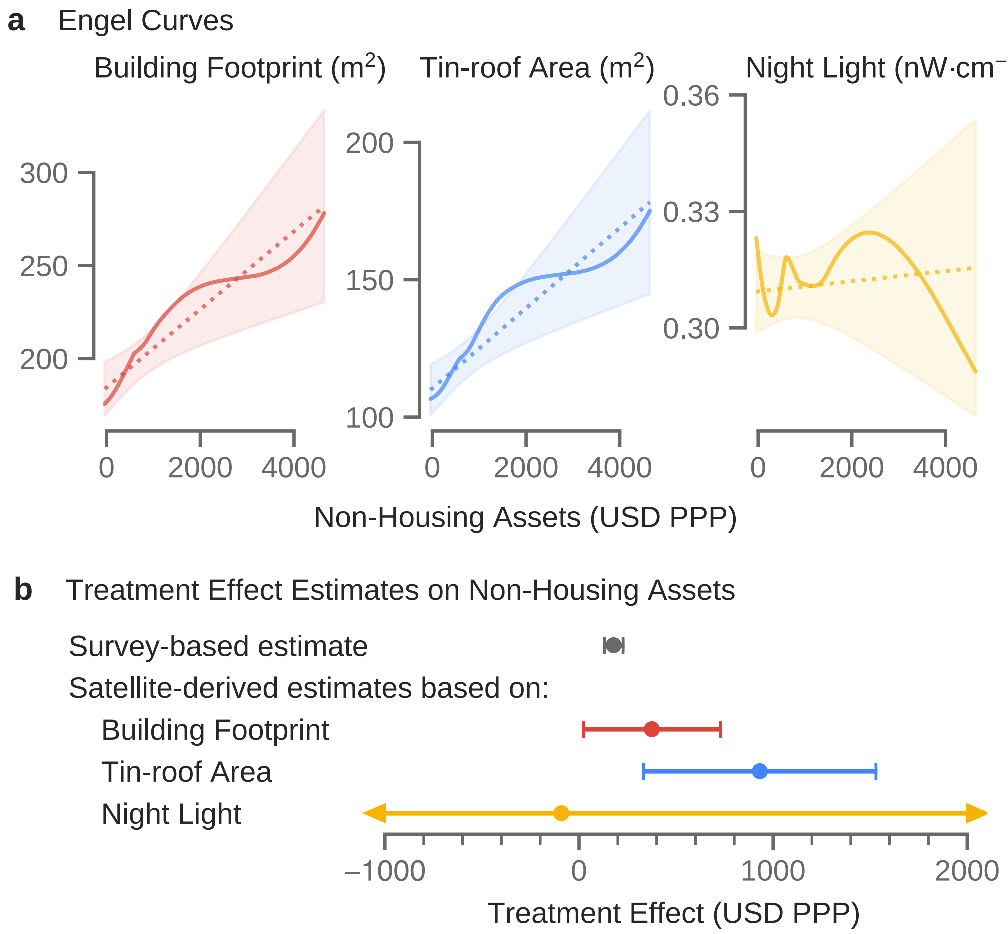

Figure S5: **The treatment effect on non-housing assets can be similarly recovered by scaling the effect on building footprint. a** The Engel curves of building footprint, tin-roof area, and night light, estimated with LOESS (solid line) or a linear regression (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the survey-based versus satellite-derived treatment effects. The dots show the point estimates. The error bars show the 95% confidence intervals, with the arrow(s) marking upper/lower bounds that are out of range (if any). $n = 1,844$.

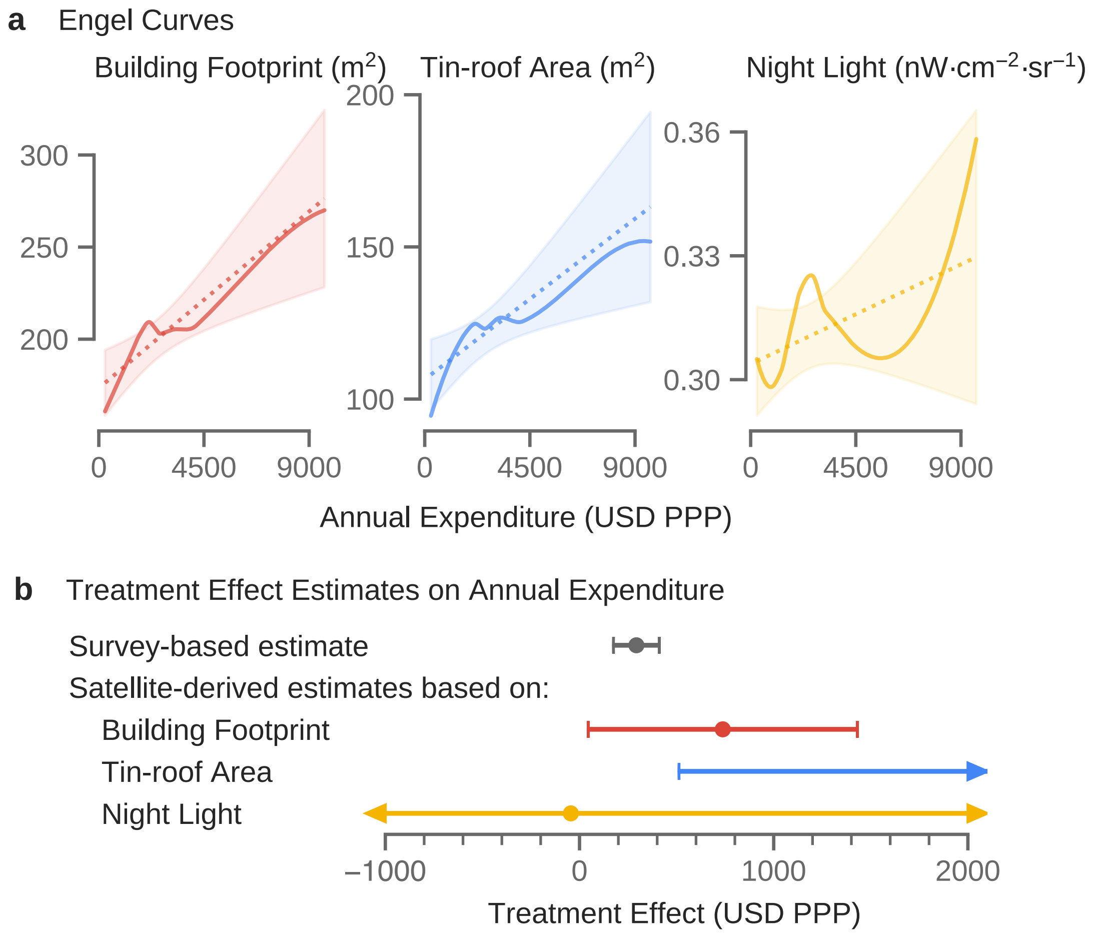

Figure S6: **The treatment effect on annual expenditure can be similarly recovered by scaling the effect on building footprint. a** The Engel curves of building footprint, tin-roof area, and night light, estimated with LOESS (solid line) or a linear regression (dotted line). The shaded regions represent the 95% confidence intervals for the latter. **b** Comparing the survey-based versus satellite-derived treatment effects. The dots show the point estimates. The error bars show the 95% confidence intervals, with the arrow(s) marking upper/lower bounds that are out of range (if any). $n = 1,843$.

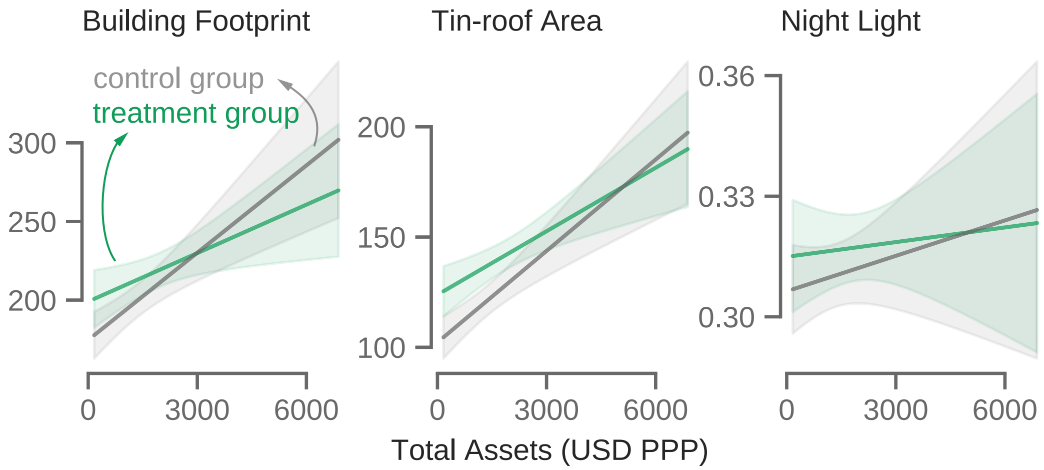

Figure S7: **The Engel curves for tin-roof area shifted in response to the cash transfer.** The Engel curves for the treatment households (in green, $n = 1,904$) and the control households (in gray, $n = 1,844$). The shaded regions represent the 95% confidence intervals.

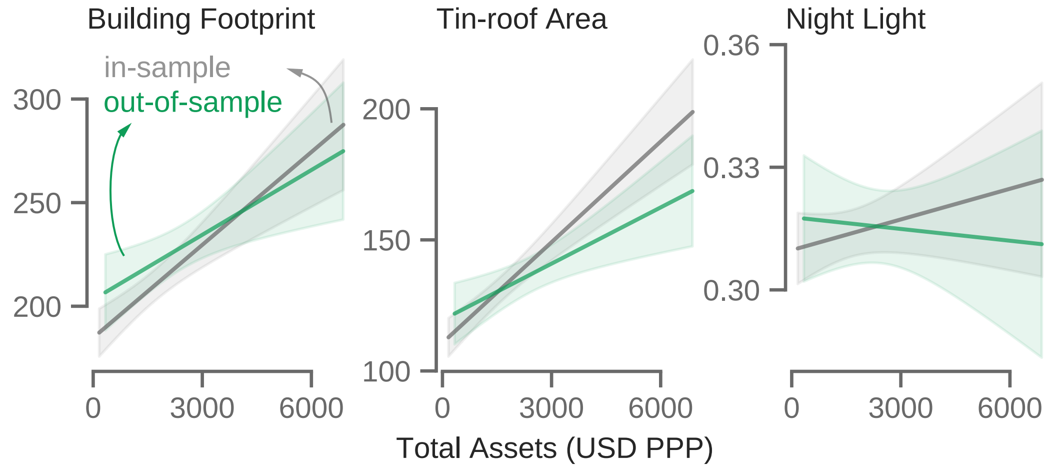

Figure S8: **The Engel curves estimated based on in-sample and out-of-sample data are broadly similar.** The Engel curves for the in-sample eligible households (in gray, $n = 3,748$) and the out-of-sample ineligible households (in green, $n = 1,821$) in the GiveDirectly study area. The shaded regions represent the 95% confidence intervals.

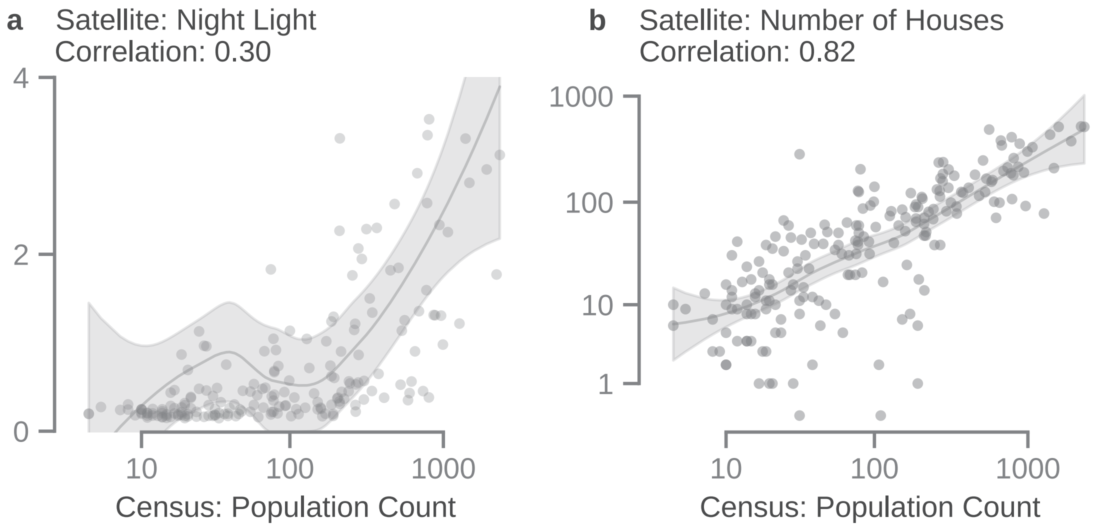

Figure S9: **Population count in Mexican villages is more strongly correlated with the number of houses in satellite imagery, compared to night light.** The population count is shown in log scale. Each point corresponds to a randomly sampled rural locality in Mexico. Gray lines are estimated LOESS curves, and the shaded regions are the 95% confidence intervals. The (Pearson) correlation coefficients are reported in the panel subtitles. $n = 197$.