Stanford | King Center on Global Development

Using Satellite Imagery to Understand and Promote Sustainable Development

Marshall Burke

Anne Driscoll

David Lobell

Stefano Ermon

October, 2020

Working Paper No. 1078

NBER WORKING PAPER SERIES

USING SATELLITE IMAGERY TO UNDERSTAND AND PROMOTE SUSTAINABLE DEVELOPMENT

Marshall Burke
Anne Driscoll
David Lobell
Stefano Ermon

Working Paper 27879
http://www.nber.org/papers/w27879

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2020

We thank Jenny Xue, Brian Lin, and Zhongyi Tang for excellent research assistance, and thank USAID Bureau for Food Security, the Global Innovation Fund, Darpa World Modelers program, and the Stanford King Center on Global Development for funding. Data and code for replication of all results will be made public upon publication. M.B., D.L., and S.E. are co-founders of AtlasAI, a company that uses machine learning to measure economic outcomes in the developing world. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2020 by Marshall Burke, Anne Driscoll, David Lobell, and Stefano Ermon. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Using Satellite Imagery to Understand and Promote Sustainable Development
Marshall Burke, Anne Driscoll, David Lobell, and Stefano Ermon
NBER Working Paper No. 27879
October 2020
JEL No. C45,C55,O1

ABSTRACT

Accurate and comprehensive measurements of a range of sustainable development outcomes are fundamental inputs into both research and policy. We synthesize the growing literature that uses satellite imagery to understand these outcomes, with a focus on approaches that combine imagery with machine learning. We quantify the paucity of ground data on key human-related outcomes and the growing abundance and resolution (spatial, temporal, and spectral) of satellite imagery. We then review recent machine learning approaches to model-building in the context of scarce and noisy training data, highlighting how this noise often leads to incorrect assessment of models' predictive performance. We quantify recent model performance across multiple sustainable development domains, discuss research and policy applications, explore constraints to future progress, and highlight key research directions for the field.

Marshall Burke
Department of Earth System Science
Stanford University
Stanford, CA 94305
and NBER
mburke@stanford.edu

Anne Driscoll
Stanford University
616 Serra St
Stanford, CA 94305
anne.driscoll@stanford.edu

David Lobell
Stanford University
Department of Environmental
Earth System Science
Y2E2 Bldg - MC4205
473 Via Ortega, room 367
Stanford, CA 94305
dlobell@stanford.edu

Stefano Ermon
Stanford University
353 Serra Mall
Stanford, CA 94305
ermon@cs.stanford.edu

1 Introduction

Humans have long sought to image their habitat from above the ground. Socrates purportedly stated in 500 B.C.E. that "Man must rise above the earth – to the top of the atmosphere and beyond – for only thus will he fully understand the world in which he lives".[1] His lofty goal was taken up in earnest after the advent of photography in the mid-nineteenth century C.E., with earth observation data collected by strapping cameras to balloons, kites, and pigeons. The first known image of earth from space was taken nearly a century later (1946) by American scientists using a captured Nazi rocket, revealing blurry expanses of the American Southwest.[2] This was followed decades later by the launch of the first civilian earth-observing satellite, Landsat I, in 1972, which ushered in the modern era of satellite-based remote sensing. As of early 2020, there are an estimated 713 active non-military earth observation satellites in orbit, 75% of which were launched in the last five years.[3] These satellites are now capturing imagery of the earth in unprecedented temporal, spatial, and spectral frequency.

Here we review and synthesize a rapidly growing scientific literature that seeks to use this satellite imagery to measure and understand various human outcomes, including a range of outcomes directly linked to the Sustainable Development Goals. We pay particular attention to recent approaches that use methods from artificial intelligence to extract information from images, as these methods typically outperform earlier approaches, enabling new insights. Our focus is on settings and applications where humans themselves, or what they produce, are the outcome of interest, and where these outcomes are being predicted using satellite imagery. We quantify existing performance in these domains across a large set of studies, explore key constraints to future progress, and highlight a number of research directions that we believe are key if these approaches are going to be improved and adopted by practitioners.

We do not review and assess the large literature on using remote sensing for other earth observation tasks (e.g. environmental monitoring), or efforts that use other sources of non-traditional, unstructured data (e.g. data from social media or cell phones) to measure human-related outcomes. We discuss this work if these other unstructured data sources are used in combination with imagery for sustainability tasks. Our review complements existing sector specific reviews, including the use of remote sensing in agriculture,[4,5] in economic applications,[6] and in the detection of informal settlements,[7] drawing common lessons across these and other domains.

Our review makes four main points. First, satellite-based performance in predicting key sustainable development outcomes is reasonably strong and appears to be improving. Indeed, analyses suggest that reported model performance likely understates true performance in many settings, given

the noisy data on which predictions are evaluated and the types of noise typically observed in sustainability applications. For multiple outcomes of interest, satellite-based estimates can now equal or exceed the accuracy of traditional approaches to outcome measurement.

Second, perhaps the largest constraint to model development is now training data rather than imagery. While imagery has become abundant, the scarcity and (in many settings) unreliability of quality ground data make both training and validation of satellite-based models difficult. Expanding the quantity and – in particular – the quality of labels will quickly accelerate progress in this field.

Third, despite the growing power of satellite-based approaches, we argue that in most settings, these approaches will amplify rather than replace existing ground-based data collection efforts. Many outcomes of interest will likely never be accurately estimated with satellites; for outcomes where satellites do have predictive power, high-quality local training data can nearly always improve model performance.

Finally, there remain few documented cases where satellites have been operationalized into public-sector decision-making processes in the sustainable development domains where we focus – with applications in population and agricultural measurements being the main exceptions. Limited adoption is likely driven by a number of forces, including the recency of the technology, the lack of accuracy (perceived or real) of the models, lack of model interpretability, and entrenched interests in maintaining the current data regime. We discuss how some of these constraints might be overcome.

2 The availability and reliability of data

2.1 Key data are scarce, and often scarcest in places where most needed

Household- or field-level surveys remain the main data collection tool for key development-related outcomes, including poverty, agricultural productivity, population, and many health outcomes. Methodologies for such data collection are well developed, and are implemented by national statistical agencies and other organizations in nearly all countries of the world. For livelihood surveys designed to generate regionally or nationally-representative estimates, sampling strategies typically follow two-stage designs, where survey “enumeration areas" (or “clusters", often the size of a village or a neighborhood) are first sampled proportional to population, and then a given number of households or individuals are randomly sampled within each cluster. Typically survey sizes for surveys such as the Demographic and Health Surveys (DHS) or Living Standard Measurement Surveys (LSMS) are a few hundred to a few thousand clusters, and then 10-20 households per cluster, yielding total household sample sizes typically between 2000 and 20,000 for a given country.

Such surveys provide critical information – and often incredible detail – on a range of outcomes, and are the bedrock on which many sustainable development related outcomes have and will continue to be measured. But their implementation and use also faces a number of important challenges. First, nationally-representative surveys are expensive and time-consuming to conduct. Conducting a DHS or LSMS survey in one country for one year typically costs $1.5-2 million USD,[8] with the entire survey operation taking multiple years and involving the training and deployment of enumerators to often remote and insecure locations. Population censuses are substantially more expensive, costing tens to hundreds of millions of USD in a typical African country.[9]

An implication of this expense is that many countries conduct surveys infrequently, if at all. In half of African nations at least 6.5 years pass between nationally representative livelihood surveys, as shown in Figure 1a (compare to sub-annual frequency in most wealthy countries). Globally, the frequency of these economic household surveys is on average substantially lower in less wealthy countries (Fig 1b), meaning that data on livelihood outcomes are often lacking where they are arguably the most needed. Surveys are also much less common in less democratic societies (Fig 1c), which could at least partly reflect the desire and ability of some autocrats to limit awareness of poor economic progress.[10] The frequency of agricultural and population censuses also varies widely around the world (Fig 1d,g). For instance, 25% (n = 53) of countries have gone more than 15 years since their last agricultural census, and 8% (n = 17) countries more than 15 years since their last population census. Restricting to just African nations, 34% of countries have gone more than 15 years since their last agricultural census. For both agricultural and population data, the relationship between survey recency, income, and level of democracy is less clear, perhaps reflecting the more important role of these data in developing economies.

A second challenge for many downstream applications are that surveys are typically only representative at the national or (sometimes) regional level, meaning they often cannot be used to generate accurate summary statistics at a state, county, or more local level. This represents a challenge for a range of research or policy applications that require individual or local-level information – for instance an anti-poverty program attempting to target an intervention (e.g cash transfer) to a particular group, or a research effort aimed at studying the impact of such an intervention.

Third, underlying household or cluster-level observations are not made publicly available in many surveys, including nearly all the surveys that contribute to official poverty statistics (such as those depicted in Fig 1a), and no geographic information is publicly provided on where in a country the data were collected. These factors further deepen the challenge of using such data to conduct local research or policy evaluation, or to train models to predict local outcomes using these data. Even when local-level anonymized georeferenced data are made public in some form, data are typically

released more than a year after survey completion, hampering real-time knowledge of livelihood conditions on the ground.

Finally, as explored below, ground data can have multiple sources of noise or bias, further limiting their reliability and utility in research and decision-making. This in turn has important implications for how satellite-based models trained on these data are validated and interpreted.

2.2 Existing ground data can be unreliable

Even where ground data are present, several key sources of error can limit their utility. First, most outcomes are not measured directly, but rather inferred from responses to surveys. This can introduce large amounts of both random and systematic measurement errors, for example in the case of self-reported household consumption[11] or agricultural production[12] surveys. For instance, in household consumption expenditure surveys, changes to the recall period or the list of items households are questioned about can lead to household expenditure estimates that are $>25\%$ too low relative to gold standard household diaries.[11]

Lack of reliability also extends to agricultural contexts. In recent reviews of agricultural statistical systems, the World Bank noted that the "practice of 'eye observations' or 'desk-based estimation' is commonly used by agricultural officers", leading to often-conflicting estimates of key agricultural outcomes by different government ministries, and to variation over time in published statistics that cannot easily be reconciled with events on the ground.[12] Current practices are likely to have a bias toward overestimation, further weakening the quality of food security assessments.[12,13]

An additional key source of noise comes from sampling variability. As noted, surveys are typically designed to be representative at very large scales (e.g. nationally), and this representativeness is typically obtained by taking small random samples of households or fields across many cluster locations. Because most agricultural and economic outcomes of interest often exhibit substantial variation even at very local levels (e.g. coefficients of variation > 1 at the village level), these small samples thus represent an unbiased but potentially very noisy measure of average outcomes in a given locality.

The combined effects of both measurement error and sampling variability can be appreciated when comparing two independent measures of the same outcome for the same administrative level. In Figure 2, average maize yields (in units of tons per ha of land) are compared at the first administrative level (e.g., province or state) as obtained from household surveys covered by the LSMS-ISA program versus by official government ministry estimates in three African countries. This comparison reveals both a systematic bias towards higher yields in official government data

than in household responses, and a relatively low correlation between the two measures, with the highest observed correlation equal to $r = 0.39$ for Ethiopia.

A third source of error, particularly relevant to researchers relying on access to data acquired by others, is noise purposefully introduced to protect the privacy of surveyed households. Adding jitter to village coordinates is common practice for most of the publicly released datasets based on household surveys, for instance with up to 2km of random jitter added in urban areas and 5km in rural areas. Below we explore the implications of these three sources of error for model development and evaluation.

2.3 Availability of satellite imagery changing rapidly

Information from satellite imagery has long offered a potential inroad into helping solve problems of data scarcity and unreliability in sustainability. Such information has been used in both agricultural and socioeconomic applications for decades.[14,15] However, thanks to both public and private sector investment, recent years have seen a remarkable increase in the temporal, spatial, and spectral information available from satellites. These investments have largely undone the traditional trade-off between temporal and spatial resolution, and are helping to undo the trade-off between spectral and temporal/spatial resolution.

To quantify this increase and understand how it varies across developing and developed countries, we randomly sample 100 locations in Africa and 100 additional across the US and EU (sampling proportional to population), and query the availability of cloud-free imagery (defined as $<$30% cloud cover) at each location in 2010 and 2019 for all available optical sensors, using multiple online tools (see Supplemental Information for details on this process). We calculate region- and year-specific average revisit rates as the number of available cloud-free images across locations divided by the number of locations times the number of days. We calculate this separately for each sensor and and also calculate an imagery-resolution "frontier", defined as overall revisit rate across sensors at or below a given spatial resolution.

Results are shown in Figure 3. Many new public and private-sector entrants since 2010 (Fig 3a) have lessened the traditional temporal/spatial tradeoff in imagery, particularly at resolutions $\geq$3m. Although the revisit rate of very high resolution ($<$1m) sensors over Africa has seen only slight improvement over the last decade (Fig 3b), and very-high-resolution revisit rates remain lower in Africa as compared to the US/EU (Fig 3c), revisit rates for high resolution (1-5m) and moderate-to low-resolution sensors has increased dramatically. Images at this resolution arenow captured multiple times per week rather than multiple times per year and equitable capture between Africa versus the US/EU.

Figure 3 provides additional detail and sample imagery for a number of sensors in African locations. Information on human activity is readily visible even in moderate-resolution sensors (5-30m), and indices constructed from moderate-resolution multispectral imagery provide an increasingly clear picture of a broad range of human activity at very local scale, including urban infrastructure development, agricultural activity, and moisture availability (Fig 3f). The increasingly high revisit rate of such imagery also provides key insight into development-relevant activities that change seasonally, such as the location and productivity of croplands (Fig 3g).

3 Modeling approaches using satellite imagery to predict sustainability outcomes

Researchers have taken many different modeling approaches in using this large amount of new imagery to measure and understand sustainable development. We use "model" to mean any function or set of functions mapping inputs (e.g., satellite images) to outputs (e.g., a wealth index or yield estimates for an area). Such models are often simple, such as linear regression models that relate satellite-derived vegetation indices to crop yields[16] or that relate nighttime lights to economic outcomes.[17] When there is substantial prior knowledge of the likely relationship between satellite-derived features and the outcome of interest, as in the case of many agricultural variables, such approaches can often work well. However, even in these settings, machine learning approaches that seek to more flexibly learn – rather than specify – the mapping of inputs to outputs can often improve predictive performance.

Machine learning approaches start by defining a suitable model family, i.e., a set of candidate functions used to represent the relationship between inputs and outputs. These could be decision trees, random forests, support vector machines, or fully-connected neural networks with a fixed structure and varying weights.[18] When inputs and outputs have explicit spatial or temporal structure (e.g. images, or images over time) it is typically advantageous to use functions tailored to this structure. These include convolutional neural networks for images, recurrent neural networks for sequential data, and convolutional autoencoders when both inputs and outputs have spatial structure[19](e.g., segmentation of agricultural fields).

Training data for these models consists of a set of inputs with their corresponding ground-truth outputs, e.g., images of villages and their corresponding poverty levels, or a sequence of images of a field captured during the growing season and the corresponding crop yield. A model in the family is chosen by training, which typically involves the minimizing of a suitable loss function that describes the difference between predicted and observed values of the outcome. For regression, the loss could be squared loss or absolute value, and for classification a common choice is cross-

entropy. After training, the loss function is evaluated on held-out data – i.e. data not used to train the model. Evaluation on held-out data is important because training data are often limited and the model family complex (often with many orders of magnitude more model parameters than training observations), and overfitting is thus a major concern. Regularization techniques such as weight decay, dropout, and early stopping using a validation set are often employed in practice to mitigate overfitting.

Suitable preprocessing of inputs is also often important in achieving good performance. Common pre-processing steps include median compositing across images to mitigate the effect of occlusions due to clouds, imputing missing values, scaling to put the all the inputs on the same scale, centering, whitening, and harmonic preprocessing for temporal data. While deep models can in principle learn these tranformations, these are tailored to existing learning algorithms and initialization schemes and will generally make learning more stable. Tiling and rescaling is also often necessary to match the input requirements (e.g. pixel dimensions) of exiting neural architectures of neural networks.

Here we provide an overview of the range of modeling approaches that have been used to relate satellite images to sustainable development outcomes.

Shallow models based on hand-crafted features. In some domains, prior knowledge of the physics, chemistry, or biology of the relevant processes suggest that certain functions of the inputs are likely useful for prediction. This is the case for numerous vegetation indexes (VI), which are computed from raw imagery as simple ratios of reflectances at different wavelengths and are known to be related to vegetation health. Simple regression models such as linear regression or random forests can be used to make pixel-wise predictions directly from these hand-crafted features to the outputs of interest, e.g. predicting yield with VIs observed over time (see ref[20] for a recent review in the agricultural domain). When the input has spatial structure, simple aggregation strategies can be used to map pixel-wise features to image-wise features. These include simple statistics such as taking the mean, quantiles (min,median,max), or histograms of binned values as inputs to a regression or ML model. As an example, this strategy is very effective for predicting GDP with nightlights[17] or aggregate crop yields at the county and state level from multispectral images.[21] However, these simple aggreagation strategies discard most of the spatial structure, which can be undesirable.

Models that use spatial structure in the imagery. In computer vision, spatial context can often greatly improve prediction accuracy for image prediction and analysis tasks. Machine learning models with filters designed to take into account spatial structure, such as convolutional neural networks (CNNs), often perform much better than hand-crafted features and aggreagation strategies.

Models such as VGG,[22] or deeper models with residual connections such as DenseNET or ResNet[23] are often employed. In this case, features are automatically learned from the data rather than hand-crafted. This is currently the leading approach in most computer vision applications, including in the satellite space when training data are plentiful. Use of this approach in sustainable development applications has proliferated in recent years, including in the measurement of population,[24–26] economic livelihoods,[27–30] infrastructure quality,[31,32] land use,[33,34] informal settlements,[35,36] fishing activity,[37,38] and many others.

Models that use spatial and temporal structure in the imagery. When available, multiple images of the same location over time can reduce ambiguity (e.g., due to partial cloud cover) and provide provide crucial information about changes occurring on the ground. Such a sequence of images is similar to a video, and architectures from video prediction in computer vision can be brought to bear for prediction and regression tasks. These include recurrent neural network variants such as long-short term memory networks (LSTMs),[39] convolutional LSTMs,[40] and 3-D CNNs, where images are fed in sequence into the model before it makes a prediction. These models have been successfully used for crop classification,[41–43] crop yield prediction,[21,44] predicting landslide susceptibility,[45] assessing building damage after disasters[46,47] among many other tasks.

Models that use several modalities. When multiple data modalities are available, such as measurements from different satellites, it is often possible to combine all the inputs into a single deep learning model. Approaches include stacking the inputs as additional channels of a single network, or multi-branch architectures where data modalities are processed separately to extract features which are then concatenated before a final prediction layer. Examples of this approach include models that combine multiple sources of satellite information[30] or models that combine imagery with data from weather sensors,[48] cell phones,[29] Wikipedia,[49] social media,[50] street-level imagery[51] or Open Street Map[52] to predict development-related outcomes.

3.1 Model development with limited training data

An additional set of techniques have been developed to utilize the above modeling approaches in the context of limited training data – a common problem in sustainability applications. For instance, standard convolutional neural network architectures contain millions to tens of millions of trainable parameters,[53] whereas training data for specific sustainability tasks can often number in the hundreds. This limited amount of labeled data is often insufficient for “end-to-end" training of deep networks, i.e. training a model to directly predict the outcome of interest on the available labeled data by minimizing a suitable loss function. Multiple strategies have been deployed to address this problem.

Using synthetic data. A first approach is to generate and use synthetic data to train models. In some cases, domain knowledge about the relevant physical process exists in the form of validated simulators. These simulators can be used to provide synthetic training data, i.e., synthetic inputs of what the process would look like from space paired with simulated outputs. These synthetic pairs can be used to augment the training data. For example, crop model simulations have been used to augment field data collection for satellite-based yield mapping in smallholder systems, and have been shown to perform on par or better than approaches that calibrate directly to limited field data.[16,54]

Transfer learning. A second approach, transfer learning, is a common strategy in deep learning. The idea is that a neural network can be pre-trained on a different but related task for which large amounts of labeled data are available (such as ImageNet in computer vision, or Functional Map of the World[55] and WikiSatNet[56] for satellite images). The neural network is then "fine-tuned" on the task of interest. For example, Jean et al[27] showed how transfer learning could use be used to predict a very small ($<$500) number of observations of economic livelihoods in Africa from imagery. A neural network was first trained to predict nightlights (a plentiful proxy for economic development) from daytime imagery, thus learning to recognize features in the high-resolution daytime imagery related to economic activity. Features were then extracted for daytime images in locations where livelihoods data were available, and a simpler model (e.g. regularized regression such as ridge or lasso) used to predict livelihoods from these features. Another recent approach applied a trained object identifier to high resolution data to identify buildings, vehicles, and other objects, and then used these objects as features in a regularized regression to predict economic wellbeing in Uganda with high accuracy.[57]

Transfer learning can also be done spatially, with models trained using data from one region where labels are often plentiful, and then "fine-tuned" on the target geography of interest where labels are sparse. To be successful, this approach requires relevant features to be similar between training and target geographies, but does not require the mapping of features to outcomes to be the same between regions (e.g. having productive crops near your house could signify wealth in one region but relative poverty in another). For example, a model trained to predict infrastructure quality in Africa could be finetuned to a specific country using only a small amount of labeled data.[32] The main challenge with spatial transfer learning is that changes in the input data distribution from one region to another (e.g. the appearance of houses or crops) will decrease predictive performance.

Unsupervised or semi-supervised learning. A third approach uses unsupervised or semi-supervised learning, which take advantage of the fact that while labels are often scarce in sustainability applications, obtaining large amounts of unlabeled satellite imagery is relatively easy.

Utilizing large amounts of unlabeled data to pre-train neural networks and learn useful features has recently shown great progress in computer vision,[58,59] narrowing the gap with fully supervised methods. Among others,[60] Tile2Vec is an unsupervised pre-training technique tailored specifically to satellite images that performs well on a range of tasks, such as crop type classification and predicting economic wellbeing in Africa.[61] Semi-supervised learning strategies attempt to improve model performance by additionally leveraging a small amount of labeled data. These are often based on the assumption that data is clustered, and decision boundaries should separate these clusters as much as possible. This idea has been extended to regression problems, with resulting performance improvements in predicting economic well-being from satellite imagery.[62]

3.2 Model development and evaluation with noisy data

The performance of satellite-based models, particularly in settings beyond where they were trained, is perhaps the most common and important concern for researchers and policy makers interested in potential applications in sustainable development. Noisy training data can degrade model performance in two ways. First, it can diminish the ability of a model to learn features in imagery that are predictive of the outcome of interest. Second, and more subtly, the model might learn relevant features but perform poorly in predicting test data, precisely because the test data has noise. This latter outcome would lead researchers to understate the model's true performance. As noisy datasets are increasingly employed for model development, researchers must contend with the dual challenges of not overfitting to noise and not underestimating the performance of a model with respect to reality. Both challenges are potentially important, with existing work mainly highlighting how noise in training data can degrade model performance.[63] But in many sustainable development settings, we believe models can learn to separate signal from noise in training data, and that the more fundamental – and underappreciated – challenge is in accurately assessing model performance in light of noisy test data. We quantify this insight and discuss methods for addressing it.

Noisy training versus noisy test data Studies in the broader computer vision/deep learning domain have demonstrated how models trained on noisy but numerate labels can still perform well when evaluated on high-quality test data, even when high-quality labels are massively outnumbered by low-quality labels in training data.[64–66] Under suitable assumptions on the noise, these empirical results can be explained from a theoretical point of view.[66,67] In sustainable development settings, while noisy training can certainly still degrade model performance when the amount of training data is limited (ref[68] provides one example in Indian smallholder wheat systems) or errors non-random (as in the poor-quality government data in Fig 2), numerous recent studies highlight how such noise can be overcome so long as training data are reasonably numerous and errors are

largely random. For instance, in Uganda, a model trained to predict maize yields from relatively noise data performed twice as well when evaluated on high-quality test data as when evaluated on noisy data held-out from training.[54] In India, a satellite-based crop classification model trained on labels derived from millions of imperfectly geolocated smartphone photos was able to exceed the performance of benchmark satellite-based classifiers.[69] A global study showed how noisy object labels from Open Street Map could be used to train a model to make accurate predictions of the location of urban structures.[70]

Using data and imagery from an earlier study of asset wealth across thousands of African villages,[30] we use simulation to explore the influence on model performance of three types of error common in publicly-available training data: (1) noise due random noise ("jitter") purposely added to village geo-coordinates to protect respondent privacy, (2) sampling variability noise from the construction of village-level estimates from small numbers of respondent households, and (3) noise from households' misreporting of asset ownership. We add a given type of noise to the observed wealth estimates, train a random forest model to predict these labels from nightlights imagery on 5 folds of the data, and evaluate performance on the remaining test data that has either been similarly degraded or unaltered; we use nightlights and random forest rather than a CNN and/or optical imagery to make these experiments tractable.

As shown in Fig 4a-c, when evaluated on noisy training data, model performance degrades as increasing amounts of each type of noise is added. However, when models trained on increasingly noisy data are evaluated on un-degraded test data, model performance remains highly stable, even for large amounts of training noise. This holds true for all three common types of training data noise we explore, again suggesting that ML models can be surprisingly robust to various types of training noise.

Accurately assessing model performance. Most existing work has focused on techniques to avoid overstating model performance, including strategies discussed above to avoid overfitting during training, and the typical practice of testing models on held-out data. Here we discuss two strategies for dealing with the opposite problem: understating model performance due to noise in test data.

A first approach is to ensure that a small amount of very high-quality ground data is available for model testing. Often this can require additional investment in data collection. In using these data, the typical practice of splitting a dataset from a single heritage into training, validation, and test sets is then replaced by a practice with two different measurement approaches for training and validation on the one hand, and testing on the other – with the high-quality data reserved for testing. Typically, the data volumes needed for testing are far fewer than for training, and thus the expenses

associated with obtaining “gold-standard” measures for testing are more likely tractable.

A second strategy, particularly useful if ground data are unavailable, is to identify a variable that previous work has identified as being associated with the outcome of interest, such as weather in the case of economic output, or fertilizer in the case of agricultural productivity. This strength of association between this variable and model predictions, as measured for instance by correlation, can then be compared to the association between the variable and the (noisy) training data for the model. Because these third variables (e.g. weather) are often readily available for most locations in the world, this approach should have broad applicability.

To illustrate both of these strategies, Figure 4d-f draws on a recent study of maize yields in Uganda.[54] The left panel shows the agreement between satellite-based yield estimates and the data on which the model was trained. In this case, the training data comprised 8mx8m crop cuts (i.e. harvests from small, randomly-selected portions of a field) from 125 different maize fields in the region. Although crop-cuts are low-error measurements of productivity for the portions of a field they sample, we consider this noisy training data because of high heterogeneity within fields and the potential spatial mismatch between the crop-cut location and the satellite pixels (which are 10x10m and not perfectly aligned with the crop cut). As judged by the training fit, the model has a relatively modest explanatory power ($r^2 = 0.25$, Fig 4d). Yet the model performance is much better when predictions are compared to the “gold-standard” measure of full plot harvests, which were available for a smaller number of randomly-selected fields (Fig 4e). Similarly, the correlation between satellite estimates and self-reported fertilizer or objective measures of soil quality were the same as the correlation between crop cut yields and these measures, suggesting the “signal" in the satellite measures was as strong as that from the ground measure (Fig 4f). A similar finding was obtained in Kenya when pitting satellite-estimated maize yields against self-reported yield data.[16]

Another example of both strategies is given in ref,[30] where estimates of wealth from satellites and from ground data are each compared against independent wealth measures from census data (considered high quality) and against a measure of annual temperature, which has been shown to correlate strongly to economic outcomes. Ground data and model predictions showed similar correlation against the independent wealth measure, and both uncovered similar non-linear relationships between temperature and wealth, suggesting that the satellite-based wealth measure was roughly as trustworthy as the original ground data.

4 Applications

Researchers are actively evaluating the usefulness of satellite imagery for a range of sustainable development applications, with more work thus far focused on whether satellites can be used

to make reliable measurements of key variables of interest and comparatively less devoted to using derived measures for downstream research tasks or policy decisions. Rather than try to provide a comprehensive survey of all applications of satellite-based remote sensing in sustainable development, we focus on four domains where recent work on satellite-based measurement has been particularly active and where comparable quantitative results exist across studies. Our goal is to provide rough performance benchmarks across these domains and, where possible, diagnose constraints to further improvement. In making these comparisons, we included all published or posted (e.g. on arxiv) studies where the test statistic of interest could be obtained for the outcome of interest in a developing-world geography.

We then review the more limited set of cases where these and other satellite-based measurements have been used for research or policy tasks. Our focus is again on domains directly involving human activity, and does not encompass progress in all realms of earth or environmental observation.

Smallholder agriculture. Roughly 2.5 billion individuals, and over half of the worlds poor, are estimated to live in "smallholder" households that primarily depend on farming small plots of land for their livelihoods.[71] While remote sensing has been used in agricultural applications for decades, coarse sensor resolutions and a paucity of training data had until recently largely precluded its application in smallholder agriculture, where field sizes are often $<$0.1ha (or roughly 1 30m Landsat pixel).

Here we assemble data from recent studies attempting to predict yield at the field scale in heterogeneous smallholder environments (ref[72] provide a nice overview of yield prediction performance at more aggregate scales). Field-scale yield prediction is useful for a range of development applications, including the targeting and evaluation of agricultural interventions and the rapid monitoring of rural livelihoods. We found 11 published studies that reported comparable performance metrics for field-scale yield prediction on smallholder fields, spanning multiple continents and seven crops. All studies used relatively simple models to relate hand-crafted features (typically, vegetation indices constructed from ratios of reflectances in the visible and near-infrared wavelengths) to ground-measured yields, and nearly all evaluated models on training rather than held-out test data. While predictive performance differed widely across and within crops (Fig 5a), likely due to the enormous temporal and spatial heterogeneity present in smallholder agriculture, re-analysis of multiple studies for which replication data were available allowed insight into the determinants of model performance.

First, models trained and evaluated on more "objective" ground data – i.e. harvest data collected from crop cuts or full plot harvests – performed on average substantially better than models trained

on farmer self-reported data (Fig 5b). This finding again highlights the importance of ground-based measurement error in training and evaluating remote sensing models. Second, in settings where average field sizes were small, model performance was much higher on larger fields (Fig 5c). This difference is likely because for certain sources of error, e.g. error in field area measurement or in the georeferencing of field data, the same magnitude error is more consequential for smaller fields; a 10m georeferencing error is more consequential for a 10m-wide field as compared to a 100m-wide field.

Finally, because collecting high quality ground data is expensive and time consuming, we studied the extent to which additional training samples improve model performance. At very small sample sizes, additional training samples rapidly improved performance on held out test data (as measured by root mean squared error; Fig 5d), up to around 30-50 samples. Performance was largely stable beyond that, suggesting that – at least in the African settings represented here – adequate performance for yield prediction could be achieved with only a few dozen high quality training samples. See Table S1 for the full list of studies and estimates we included.

Population A second area in which satellite information has played an important role is helping generate local-level population estimates. Accurate knowledge of where people are is a critical input into an immense range of research and policy applications. Because population census are infrequent in many developing countries and fine-scale data from existing censuses are often not made public, generating fine-scale model-based estimates of settlement locations and population density has been an areas of substantial research focus for decades.

The traditional approach to generating local-level population estimates takes a “top-down" approach in which available admin-level census data is redistributed down to a finer-scale grid (1km or finer), using satellite-derived information and other covariates as input. Because population data are almost never available for training or validation at the target fine scale, one common approach uses the coarse-scale data from census to model the relationship between satellite features (e.g. nighttime lights imagery or satellite-derived estimates of land use), other ancillary data such as the location of transportation infrastructure, and census-based population estimates, and then applies the trained model to available fine-scale features.[73,74] Another approach generates a binary population mask at fine scale using estimates of building or settlement locations derived from imagery, and then applies this mask to coarse-scale census data.[75] Both approaches typically use machine learning at some step, e.g a random forest to predict coarse census data, or computer vision approaches to identify settlement locations. For either approach, predictions can only be readily evaluated at coarse scale; the fine scale gridded predictions cannot be easily validated. In the absence of clear evaluation opportunities, a consortium of data producers have built useful tools in which different

gridded estimates can be visually compared at local scale (https://popgrid.org).

As additional quantitative comparison, we study three commonly-used population rasters that used satellite data as at least one input in their production: WorldPop,[74] GHSL,[75] and LandScan.[73] We harmonize each to a consistent 1km grid and compare population estimates for grid cells with non-zero estimates across all three rasters. Estimates show modest agreement (r=0.62-0.78) when comparing across all global pixels (Figure 6), with lowest agreement between LandScan and the other rasters. Agreement was often substantially lower in the developing world. On the African continent, the average pairwise correlation between the 3 datasets across 47 African countries is $r = 0.45$, perhaps in part due to the relative paucity of census data on which to train models. Overall disagreements in African and globally could also result from differences in conceptualization of population used in dataset construction, with LandScan attempting to measure "ambient population" averaged over 24 hours and the other datasets attempting to measure population at individuals' usual residences. Agreement improves when comparisons are made at increasingly aggregate levels, with correlations approaching $r = 1.0$ when estimates are aggregated to 100km pixels.

Multiple studies have sought to further validate estimates of one or more of these datasets in settings where fine-scale population data are available. Using very-fine-scale (100m) administrative population data from Sweden available over a 25-yr period (none of which used in the creation of any of the gridded datasets), researchers found cell-wise correlations between the admin data and GHSL, WorldPop, and LandScan of $r = 0.83$, $r = 0.82$, and $r = 0.7$, respectively, with predictive performance improving slightly in later years.[76] The authors caution that performance in Sweden (where model predictions were highly correlated, see Fig 6) might not reflect performance elsewhere, given the high quality of ancillary data available in Sweden. Other studies in China and Europe found similar or higher performance of individual gridded datasets evaluated at somewhat more aggregate scales, but (as in Fig 6) found that performance was not uniform and tended to degrade at finer spatial scales.[77,78] Overall performance on this population prediction task appears roughly on par with performance predicting asset wealth described below.

Because standard approaches to generating these estimates is to disaggregate official census estimates, final estimates are unavoidably affected by any inaccuracies in the official census data – for instance due to the most recent census having occurred a decade or more prior. An alternative that does not face this problem is to train "bottom-up" models to directly predict local-level population estimates, and these approaches have shown promise in multiple settings.[9,26,79] Such approaches are beginning to be incorporated into global gridded products (e.g. WorldPop) for countries where censuses are particularly out of date,[80] and have been shown to be a cost-effective way for generating reliable national-scale population estimates.[9]

Economic livelihoods

Predicting variation in local-level economic outcomes is another domain where the combination of machine learning and satellite imagery has seen recent application, again motivated by the paucity of existing data (Fig 1) and the broad range of applications for which such data could be useful. As in the agricultural setting, existing work spans diverse geographies and seeks to predict a range of outcomes, making quantitative comparison of different models or sensors difficult.

We focus on 12 studies that used imagery – either alone or in combination with other data – to predict asset wealth at local level in the developing world. Asset wealth is a commonly used measure of households' longer-run economic wellbeing, and is consistently measured in a number of georeferenced nationally-representative household surveys, making it appealing training data in this domain. Fig 6a shows 16 asset wealth estimates across these 11 studies. All studies applied convolutional neural networks to imagery to generate features used to predict wealth, and reported evaluation statistics on held-out test data.

While study intercomparison was challenging even for this group of studies that measured the same outcome due to the varied geographic settings (spanning Africa, Asia, and the Caribbean), the various spatial scales at which predictions were evaluated (from village level to district level), and some studies' inclusion of additional data input data not from satellites, results allowed some generalizations. First, information derived from satellites could always explain more than half, and often more than 75%, of the variation in the survey-measured asset wealth, with performance appearing to trend upward over time. For reasons described above, these estimates likely understate true model performance, as test data are almost always from publicly-available survey data with known sources of noise. Second, although small samples make generalization tenuous, studies that made predictions at more aggregate spatial scales, and studies that combined satellite information with data from other sources, tended to outperform village-level satellite-only models. These data fusion approaches have become increasingly common, with researchers demonstrating how combining imagery with data from cell phones,[29] Wikipedia,[49] social media,[50] or Open Street Map[52] can improve predictions.

Table S2 describes results from additional studies that looked at other measures of economic livelihoods, including consumption expenditure and multi-dimensional poverty indices. Prediction performance for consumption expenditure (the measure on which official poverty estimates are based) is typically lower than that for asset wealth, a difference which has been in part attributed to relatively higher noise in the consumption data[27,30] and the extreme paucity of public georeferenced public data on which to train models.

Informal settlements A final related area where there has been much recent work is in the detection of informal settlements (sometimes called “slums"). Urban populations are growing rapidly throughout much of the developing world, and about 30% of developing-country urban populations are estimated to live in slums – settled areas where inhabitants lack access to essential services, durable housing, and/or tenure security.[81] Systematic data on the location and size of such settlements is lacking, making it difficult to monitor and target service delivery and to protect residents against eviction, among other challenges.[7] Some governments, lacking reliable data on informal settlements, do not officially acknowledge their existence.[81]

Because the spatial structure (e.g density, size and type of buildings) can differ substantially between informal settlements and surrounding regions, researchers have sought to use imagery to measure the location and size of these settlements (see ref[7] for a recent review). We focus on 23 studies that used satellite imagery to segment or classify informal settlements in the developing world. These studies use a variety of methods, with some focused on creating rule bases for classification and others on directly using machine learning for classification. The fuzzy-logic rule bases are sometimes generated using machine learning (eg. decision trees) and sometimes are human generated from ontologies (formalized descriptions of expert knowledge from a certain perspective) of local informal settlements.

As with the other domains discussed, the literature spans diverse geogrophies where informal settlements can be very structurally dissimilar from each other, making study intercomparison difficult. However, in 17 studies that reported classification accuracy (evaluated against typically small numbers of ground observations), accuracy exceeded 80% in most studies and appeared to be improving over time (Fig 6g). Table S3 shows results from additional studies that reported alternate performance metrics.

4.1 Application in research

Here we highlight a number of settings in which measures derived from satellite-based remote sensing, including those discussed above, are being used for some downstream research task in the developing world.

The widest adoption of satellite-derived measures in research and policy has been in the realm of population estimates, with existing gridded population data being used in an impressive array of research applications. These include in public health, disaster response, economic development, climate change research, and others; see refs[9,80,82] for excellent recent reviews.

Satellite imagery has also been widely used to better understand agricultural productivity, including

why some fields or some regions are more productive than others[5] and whether particular management practices have been adopted.[83] Satellite estimates are also increasingly being used to identify fields most likely to respond to a particular input[16,54] or new management practice.[84]

Fisheries and animal production are additional food-related domains where satellite imagery is becoming increasingly used in research and policy. Recent work shows how multiple satellite sensors and deep learning can shed light on overall patterns of global fishing activity[37] as well as on specific activities like illegal fishing.[38,85]

Researchers in economics also increasingly utilize satellite imagery – and particularly night-time lights imagery – for a variety of applications (see ref[6] for a review). Nightlights have been used to assess the validity of official government statistics,[17,86] to understand the growth and activity of urban versus rural areas,[87,88] and to assess the role of local and federal institutions, transport costs, and other factors on economic development.[89–92] While the use of optical imagery beyond nightlights remains somewhat more limited, recent papers have shown how high-resolution optical imagery can be used to measure compliance with conservation programs[93] and to understand how ethnic favoritism shapes economic investment.[94] Recent work[95] also shows how to combine satellite-derived estimates with survey data to obtain tighter confidence intervals and improve regression analyses.

More recent work has shown how satellites can be useful in the experimental evaluation of interventions in both the agricultural and economic sphere. Jain et al[84] show how remote sensing estimates can be used to measure the effectiveness of a new agricultural technology on productivity and quantify who benefits most from the adoption of the technology. Huang[96] shows how a deep learning model trained to identify housing quality in high-resolution imagery can be used to estimate the livelihood impact of a randomized cash transfer program in Kenya, with estimates benchmarked against ground survey data. Jayachandran et al[93] show how high-resolution imagery can be used to measure compliance in an experimental evaluation of a payment-for-forest-protection program. While all of these studies focus on settings where changes induced by an intervention are readily apparent in imagery – an aspect that might not hold in other settings – they demonstrate the large potential for satellite imagery to contribute to the quantitative evaluation of many development interventions.

4.2 Use in decision making

While satellite-based measures are now being used in a variety of research applications, documented examples of their operational use in public-sector decision-making and policy in the developing world is much more limited. Systematic information on operational use in the private sector is

even more sparse, although use is likely widespread and growing; the same is true of military applications. Here we only consider public-sector non-military use.

As in research, the widest application of satellite-based measures in public-sector decision-making is in the population domain. For instance, the UN World Food Programme and US government both used gridded population estimates to inform needs assessments and target humanitarian response following natural disasters.[80] Gridded population data are also being used to inform sampling strategies for ground surveys.[80]

In agriculture, remote-sensed vegetation indices and satellite-derived rainfall estimates are key inputs into short-term forecasting of food insecurity, which directly informs food aid and other humanitarian resource allocation.[97] Numerous systems that track agricultural growing conditions and crop output around the world also make ample use of remote sensing information, and output from these systems are used in a wide array of tasks, including in early warning alerts, foreign aid decisions, analysis of commercial trends, and in trade policy.[98] Data from remote detection of fishing activity is also being used by numerous governments and other organizations to manage fisheries and design protected areas.[99]

Across other domains – e.g. economic livelihood measurement – documented use in decision-making appears limited or non-existent, although anecdotally there is rapidly growing interest in the policy community in exploring these measures.[100]

We hypothesize on why adoption in these and other domains has been relatively limited. The simplest explanation is that the combination of satellite information and machine learning is still quite new in many domains, and decision-makers might not be familiar with these approaches or convinced they are “good enough". Our view is that in many settings, including smallholder agricultural and livelihood measurement, the true accuracy of satellite-derived estimates can rival or exceed that of traditional survey-based measures. It remains the job of the research community to help make this clear, and the job of the user community to transparently define the counterfactual: if not satellite-based data, what alternative data would be used to make a decision, and what do we know about its reliability?

Even if satellite-based measures are accurate, they might not yet be operational. To our knowledge there exist no updated, global-scale estimates of smallholder crop productivity, economic well-being, or informal settlements that a decision-maker could immediately use (estimates are beginning to exist for individual countries). The research community is arguably not well positioned to generate and update such estimates over time, and partnerships with public-sector institutions or the private sector to scale and operationalize these estimates could be important in enabling their

sustained use.

Even when models are operational, decision-makers might be understandably hesitant to adopt a measure they cannot fully explain. Deep learning models tend to sacrifice interpretability for predictive performance, and researchers are often satisfied if a model is working well (as evaluated on held-out data) even if they cannot explain why. But understanding why a model makes the predictions it does can help build trust that predictions are accurate and fair. Well-publicized instances of algorithmic bias in other settings (e.g. predictive policing, sentencing, and hiring decisions[101]), and concerns by civil rights groups that further deployment of algorithmic decision-making might worsen racial and socioeconomic inequalities,[102,103] understandably amplify worries that predictions from these new approaches could be either inaccurate or unfair.

Existing guidelines for Fairness, Accountability, and Transparency in Machine Learning ("FAT ML"),[104] if followed, could help navigate these issues. The guidelines aim to ensure that researchers are aware of potential discriminatory impacts of their algorithms and are able to investigate and provide redress should issues arise. While implementation of the guidelines certainly has its own challenges[105] (e.g. defining "fairness"), we are not aware of any of the papers we review above – including our own – having fully engaged with these guidelines.

A final reason for limited adoption is that some actors might see benefit in not having certain outcomes be measured. Autocratic regimes already collect less data (recall Fig 1), and certain countries have passed laws (since reversed) that make it a crime to publish independent estimates of key economic outcomes.[106]

5 Conclusions and directions for future work

We draw four main conclusions from the above analysis, and lay out open challenges and directions for future work. First, satellite-based performance in predicting key sustainable development outcomes is reasonably strong and appears to be improving. Estimates are being used in a wide variety of research applications and, in some cases, are already actively informing decision-making. Indeed, analyses suggest that reported model performance likely understates true performance in many settings, given the noisy data on which predictions are evaluated, and that satellite-based estimates can equal or exceed the accuracy of traditional approaches to measuring key outcomes. For certain outcomes, satellite-based approaches can already add substantial information at broad scale and low cost compared to what can be collected on the ground. Numerous quantitative approaches now exist to assist researchers and practitioners in better understanding – and not underestimating – the performance of satellite-based approaches relative to traditional alternatives.

Second, perhaps the largest constraint to model development is now training data rather than imagery. While imagery has become abundant, the scarcity and (in many settings) unreliability of quality labels make both training and validation of satellite-based models difficult. Expanding the quantity and – in particular – the quality of labels will quickly accelerate progress in this field, and allow both researchers and practitioners to measure new outcomes and to accurately assess model performance.

Third, despite the growing power of satellite-based approaches, there are many domains where such approaches are likely to contribute little in the near term – for instance, in measuring female empowerment, educational outcomes, or conflict events. Even in settings where satellites are likely to be useful, satellite-based approaches will likely amplify rather than replace existing ground-based data collection efforts. High-quality local training data can nearly always improve model performance, and will remain essential for convincing both researchers and decision-makers that satellite-based approaches are working.

Finally, there remain limited documented cases where satellites have been operationalized into decision-making processes in the sustainable development domains where we focus – with satellite-informed population estimates being the main exception. Limited adoption is likely driven by a number of forces, including the recency of the technology, the lack of accuracy (perceived or real) of the models, lack of model interpretability, and entrenched interests in maintaining the current data regime.

Helping to overcome these constraints constitute key tasks for researchers and policymaker going forward. We suggest nine specific areas where we believe future work would be particularly useful:

1. *More accurate, more numerous training data*. Many applications of deep learning outside sustainable development have been advanced by the curation of reference datasets that are then made available to the community. These datasets lower the barriers to entry and make comparison of different approaches more straightforward, yet they are lacking for sustainable development outcomes. Particularly needed are datasets that track outcomes over time so that models can be optimized to detect changes. These datasets are a major public good and investment in their collection would greatly accelerate research progress. Collecting and publishing location data from existing and ongoing ground surveys (using appropriate privacy safeguards already widely in use) would also greatly benefit research efforts in this area.
2. *More evaluation in the context of specific use cases*. Most evaluation of satellite estimates have focused on agreement with a ground-based measure of a particular outcome. Fewer

studies have then gone the next step to evaluate the actual application of the outcome measure, such as to test the impact of a randomized control trial or target an intervention to a sub-population. These downstream tasks often provide a more tangible example of the utility to potential users, and can avoid the pitfalls of direct comparisons to noisy ground measures. A related task will be to define and utilize meaningful loss functions for the specific task at hand; for instance, a poverty targeting application might be more tolerant of small errors at the wealthy end of the distribution than the poorer end.

3. *Improved model interpretability and transparency*. Especially in cases where satellite-based prediction is being used to make decisions that directly impact people (e.g. targeting aid) it is especially important that predictions be explainable and that decisions based on those predictions be transparent. Applying FAT ML or similar guidelines to research output will be increasingly important as research gets operationalized.
4. *Creative data fusion*. Combining information from multiple different optical sensors of different temporal and spatial resolutions, combining different types of imagery (e.g. optical + radar), and/or combining satellite imagery with other relevant data (e.g. from cell phones), appear to be particularly promising approaches to improving model performance. As much of these additional data are collected by the private sector, sustained and enforceable data-sharing agreements between companies and researchers will be key.[107]
5. *Scaling estimates*. Researchers typically have more incentive to innovate on methods than they do (e.g.) to apply validated methods across large geographies and update estimates as new data come in – the later being what is often required to make outputs useful to decision-makers. Partnerships between academic researchers and public- or private-sector organizations who have the skills and resources to do this scaling will be key to operationalizing many promising research advances in the satellite/ML domain.
6. *Measuring changes over time*. Much of the literature reviewed above makes predictions at a given point in time. However, many applications require measuring changes over time. While the relationship between inputs and outputs over time is reasonably stable in some domains (e.g. vegetation indices and yields in agriculture), this might not be true in other domains (e.g. economic development). Unfortunately, temporal evaluation at a local level is difficult because there exist few ground datasets that repeatedly and reliably measure the same locations over time. Curating these datasets and using them to develop and validate temporal predictions will be key for tracking the evolution of key sustainability outcomes.
7. *Using imagery to actively guide ground data collection*. As predictive performance of satellite-based models improve, their output could be used to optimally guide further data collection on the ground – for instance, to collect data in locations where model predictions are least certain. Research should explore to what extent such sampling strategies could

improve outcome measurement compared to traditional sampling approaches.

8. *Understanding potential pitfalls in causal inference applications*. For instance, can poverty predictions from a satellite-based model be used to study the impact of new road construction on poverty, if there is a chance that the model looks for a road to decide whether a location is poor? How do we proceed if we're concerned that image-derived proxies for a dependent variable of interest are themselves the independent variable of interest?
9. *Improved guidelines for privacy*. As predictions become increasingly granular and accurate, who has access to these data? How can precisely georeferenced ground data (which is increasingly collected) be used to train or validate models without undermining privacy? Guidelines for navigating these issues are increasingly critical as models improve.

References

[1] Moore, G. K. What is a picture worth? a history of remote sensing. *Hydrological Sciences Bulletin* **24**, 477–485 (1979).

[2] Waxman, O. B. Aerial photography's surprising role in history (2018). URL `https://time.com/longform/aerial-photography-drones-history/`.

[3] Union of Concerned Scientists. The ucs satellite database (2020). URL `http://www.ucsusa.org/satellite_database`.

[4] Mulla, D. J. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering* **114**, 358–371 (2013).

[5] Lobell, D. B. The use of satellite data for crop yield gap analysis. *Field Crops Research* **143**, 56–64 (2013).

[6] Donaldson, D. & Storeygard, A. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives* **30**, 171–98 (2016).

[7] Kuffer, M., Pfeffer, K. & Sliuzas, R. Slums from space—15 years of slum mapping using remote sensing. *Remote Sensing* **8**, 455 (2016).

[8] Network, S. D. S. *Data for development: A needs assessment for SDG monitoring and statistical capacity development* (Sustainable Development Solutions Network., 2015).

[9] Wardrop, N. *et al.* Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences* **115**, 3529–3537 (2018).

[10] Devarajan, S. Africa's statistical tragedy. *Review of Income and Wealth* **59**, S9–S15 (2013).

[11] Beegle, K., De Weerdt, J., Friedman, J. & Gibson, J. Methods of household consumption measurement through surveys: Experimental results from tanzania. *Journal of Development Economics* **1**, 3–18 (2012).

[12] Carletto, C., Jolliffe, D. & Banerjee, R. From tragedy to renaissance: Improving agricultural data for better policies. *The Journal of Development Studies* **51**, 133–148 (2015).

[13] Braimoh, A. *et al.* Capacity needs assessment for improving agricultural statistics in kenya. Tech. Rep., The World Bank (2018).

[14] MacDonald, R. B. & Hall, F. G. Global crop forecasting. *Science* **208**, 670–679 (1980).

[15] Elvidge, C. D. *et al.* Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing* **18**, 1373–1379 (1997).

[16] Burke, M. & Lobell, D. B. Satellite-based assessment of yield variation and its determinants in smallholder african systems. *Proceedings of the National Academy of Sciences* **114**, 2189–2194 (2017).

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

[18] Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media, 2009).

[19] Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep learning*, vol. 1 (MIT press Cambridge, 2016).

[20] Weiss, M., Jacob, F. & Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment* **236**, 111402 (2020).

[21] You, J., Li, X., Low, M., Lobell, D. & Ermon, S. Deep gaussian process for crop yield prediction based on remote sensing data. In *AAAI*, 4559–4566 (2017).

[22] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[23] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

[24] Tiecke, T. G. *et al.* Mapping the world population one building at a time. *arXiv preprint arXiv:1712.05839* (2017).

[25] Zong, Z., Feng, J., Liu, K., Shi, H. & Li, Y. Deepdpm: Dynamic population mapping via deep neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 1294–1301 (2019).

[26] Hu, W. *et al.* Mapping missing population in rural india: A deep learning approach with satellite imagery. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 353–359 (2019).

[27] Jean, N. *et al.* Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).

[28] Head, A., Manguin, M., Tran, N. & Blumenstock, J. E. Can human development be measured with satellite imagery? In *ICTD*, 8–1 (2017).

[29] Steele, J. E. *et al.* Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface* **14**, 20160690 (2017).

[30] 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).

[31] Cadamuro, G., Muhebwa, A. & Taneja, J. Assigning a grade: Accurate measurement of road quality using satellite imagery. *arXiv preprint arXiv:1812.01699* (2018).

[32] Oshri, B. *et al.* Infrastructure quality assessment in africa using satellite imagery and deep learning. In *Proc. 24th ACM SIGKDD Conference* (2018).

[33] Albert, A., Kaur, J. & Gonzalez, M. C. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1357–1366 (2017).

[34] Helber, P., Bischke, B., Dengel, A. & Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **12**, 2217–2226 (2019).

[35] Mboga, N., Persello, C., Bergado, J. R. & Stein, A. Detection of informal settlements from vhr images using convolutional neural networks. *Remote sensing* **9**, 1106 (2017).

[36] Persello, C. & Stein, A. Deep fully convolutional networks for the detection of informal settlements in vhr images. *IEEE geoscience and remote sensing letters* **14**, 2325–2329 (2017).

[37] Kroodsma, D. A. *et al.* Tracking the global footprint of fisheries. *Science* **359**, 904–908 (2018).

[38] Park, J. *et al.* Illuminating dark fishing fleets in north korea. *Science Advances* **6**, eabb1197 (2020).

[39] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).

[40] Xingjian, S. *et al.* Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 802–810 (2015).

[41] Ji, S., Zhang, C., Xu, A., Shi, Y. & Duan, Y. 3d convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sensing* **10**, 75 (2018).

[42] Rußwurm, M. & Körner, M. Multi-temporal land cover classification with long short-term memory neural networks. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **42**, 551 (2017).

[43] M Rustowicz, R. *et al.* Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 75–82 (2019).

[44] Sun, J., Di, L., Sun, Z., Shen, Y. & Lai, Z. County-level soybean yield prediction using deep cnn-lstm model. *Sensors* **19**, 4363 (2019).

[45] Xiao, L., Zhang, Y. & Peng, G. Landslide susceptibility assessment using integrated deep learning algorithm along the china-nepal highway. *Sensors* **18**, 4436 (2018).

[46] Xu, J. Z., Lu, W., Li, Z., Khaitan, P. & Zaytseva, V. Building damage detection in satellite imagery using convolutional neural networks. *arXiv preprint arXiv:1910.06444* (2019).

[47] Ci, T., Liu, Z. & Wang, Y. Assessment of the degree of building damage caused by disaster using convolutional neural networks in combination with ordinal regression. *Remote Sensing* **11**, 2858 (2019).

[48] Davenport, F. M. *et al.* Using out-of-sample yield forecast experiments to evaluate which earth observation products best indicate end of season maize yields. *Environmental Research Letters* **14**, 124095 (2019).

[49] Sheehan, E. *et al.* Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2698–2706 (2019).

[50] Fatehkia, M., Coles, B., Ofli, F. & Weber, I. The relative value of facebook advertising data for poverty mapping. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 934–938 (2020).

[51] Cao, R. *et al.* Integrating aerial and street view images for urban land use classification. *Remote Sensing* **10**, 1553 (2018).

[52] Tingzon, I. *et al.* Mapping poverty in the philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. In *AI for Social Good ICML 2019 Workshop* (2019).

[53] Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).

[54] Lobell, D. B. *et al.* Eyes in the sky, boots on the ground: Assessing satellite-and ground-based approaches to crop yield measurement and analysis. *American Journal of Agricultural Economics* **102**, 202–219 (2020).

[55] Christie, G., Fendley, N., Wilson, J. & Mukherjee, R. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6172–6180 (2018).

[56] Uzkent, B. *et al.* Learning to interpret satellite images using wikipedia. *IJCAI* (2019).

[57] Ayush, K., Uzkent, B., Burke, M., Lobell, D. & Ermon, S. Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612* (2020).

[58] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738 (2020).

[59] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* (2020).

[60] Basu, S. *et al.* Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, 1–10 (2015).

[61] Jean, N., Wang, S., Azzari, G., Lobell, D. & Ermon, S. Tile2vec: Unsupervised representation learning for remote sensing data. In *arXiv preprint arXiv:1805.02855* (2018).

[62] Jean, N., Xie, S. M. & Ermon, S. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. In *NIPS* (2018).

[63] Elmes, A. *et al.* Accounting for training data error in machine learning applied to earth observations. *Remote Sensing* **12**, 1034 (2020).

[64] Krause, J. *et al.* The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, 301–320 (Springer, 2016).

[65] Rolnick, D., Veit, A., Belongie, S. & Shavit, N. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694* (2017).

[66] Natarajan, N., Dhillon, I. S., Ravikumar, P. K. & Tewari, A. Learning with noisy labels. In *Advances in neural information processing systems*, 1196–1204 (2013).

[67] Charikar, M., Steinhardt, J. & Valiant, G. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 47–60 (2017).

[68] Paliwal, A. & Jain, M. The accuracy of self-reported crop yield estimates and their ability to train remote sensing algorithms. *Frontiers in Sustainable Food Systems* **4**, 25 (2020).

[69] Wang, S. *et al.* Mapping crop types in southeast india with smartphone crowdsourcing and deep learning. *working paper* (2020).

[70] Kaiser, P. *et al.* Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing* **55**, 6054–6068 (2017).

[71] Christen, R. P. & Anderson, J. Segmentation of smallholder households: Meeting the range of financial needs in agricultural families. Tech. Rep., The World Bank (2013).

[72] Chivasa, W., Mutanga, O. & Biradar, C. Application of remote sensing in estimating maize grain yield in heterogeneous african agricultural landscapes: a review. *International Journal of Remote Sensing* **38**, 6816–6845 (2017).

[73] Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C. & Worley, B. A. Landscan: a global population database for estimating populations at risk. *Photogrammetric engineering and remote sensing* **66**, 849–857 (2000).

[74] Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. J. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS one* **10**, e0107042 (2015).

[75] Schiavina, M., Freire, S. & MacManus, K. Ghs population grid multitemporal (1975, 1990, 2000, 2015) r2019a. *Eur. Comm. JRC* (2019).

[76] Bustos, M. F. A., Hall, O., Niedomysl, T. & Ernstson, U. A pixel level evaluation of five multitemporal global gridded population datasets: a case study in sweden, 1990–2015. *Population and Environment* (2020).

[77] Calka, B. & Bielecka, E. Ghs-pop accuracy assessment: Poland and portugal case study. *Remote Sensing* **12**, 1105 (2020).

[78] Bai, Z., Wang, J., Wang, M., Gao, M. & Sun, J. Accuracy assessment of multi-source gridded population distribution datasets in china. *Sustainability* **10**, 1363 (2018).

[79] Engstrom, R., Newhouse, D. & Soundararajan, V. Estimating small-area population density in sri lanka using surveys and geo-spatial data. *PloS one* **15**, e0237063 (2020).

[80] Thematic Research Network on Data and Statistics. Leaving no one off the map: A guide for gridded population data for sustainable development. Tech. Rep., UN Sustainable Development Solutions Network (2020).

[81] Habitat, U. Habitat iii issue paper 22—informal settlements. *New York: UN Habitat* (2015).

[82] Leyk, S. *et al.* The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth System Science Data* **11** (2019).

[83] Kubitza, C., Krishna, V. V., Schulthess, U. & Jain, M. Estimating adoption and impacts of agricultural management practices in developing countries using satellite data. a scoping review. *Agronomy for Sustainable Development* **40**, 1–21 (2020).

[84] Jain, M. *et al.* The impact of agricultural interventions can be doubled by using satellite data. *Nature Sustainability* **2**, 931–934 (2019).

[85] Belhabib, D. *et al.* Catching industrial fishing incursions into inshore waters of africa from space. *Fish and Fisheries* **21**, 379–392 (2020).

[86] Pinkovskiy, M. & Sala-i Martin, X. Lights, camera... income! illuminating the national accounts-household surveys debate. *The Quarterly Journal of Economics* **131**, 579–631 (2016).

[87] Henderson, J. V., Squires, T., Storeygard, A. & Weil, D. The global distribution of economic activity: nature, history, and the role of trade. *The Quarterly Journal of Economics* **133**, 357–406 (2018).

[88] Harari, M. Cities in bad shape. *American Economic Review* **110**.

[89] Michalopoulos, S. & Papaioannou, E. Pre-colonial ethnic institutions and contemporary african development. *Econometrica* **81**, 113–152 (2013).

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

[91] Storeygard, A. Farther on down the road: transport costs, trade and urban growth in sub-saharan africa. *The Review of economic studies* **83**, 1263–1295 (2016).

[92] Pinkovskiy, M. L. Growth discontinuities at borders. *Journal of Economic Growth* **22**, 145–192 (2017).

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

[94] 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).

[95] Zhao, S., Yeh, C. & Ermon, S. A framework for sample efficient interval estimation with control variates. In *International Conference on Artificial Intelligence and Statistics*, 4583–4592 (2020).

[96] Huang, L. Y. Measuring the impacts of poverty alleviation programs with satellite imagery and deep learning (2020). `http://luna-yue-huang.com/assets/pdf/jmp.pdf`.

[97] Brown, M. E. *Famine early warning systems and remote sensing data* (Springer Science & Business Media, 2008).

[98] Fritz, S. *et al.* A comparison of global agricultural monitoring systems and current gaps. *Agricultural systems* **168**, 258–272 (2019).

[99] Watch, G. F. *Ocean sustainability through transparency, data-sharing and collaboration* (2020). URL `https://globalfishingwatch.org/wp-content/uploads/GFW-program-2020.pdf`.

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

[101] Cossins, D. Discriminating algorithms: 5 times ai showed prejudice. *New Scientist* **12** (2018).

[102] Milner, Y. Data for black lives. URL `https://d4bl.org/about.html`.

[103] on Civil Human Rights, L. C. Civil rights principles for the era of big data. URL `https://civilrights.org/2014/02/27/civil-rights-principles-era-big-data/`.

[104] Diakopoulos, N. *et al.* Principles for accountable algorithms and a social impact statement for algorithms. *FAT/ML* (2017).

[105] Gajane, P. & Pechenizkiy, M. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).

[106] It is no longer a crime to publish statistics in tanzania. *The Citizen* (2019). URL https://www.thecitizen.co.tz/news/It-is-no-longer-a-crime-to-publish-statistics-in-Tanzania-/1840340-5174870-wjjdxhz/index.html.

[107] Lazer, D. M. *et al.* Computational social science: Obstacles and opportunities. *Science* **369**, 1060–1062 (2020).

[108] Solt, F. The standardized world income inequality database, version 8. *Cambridge: Harvard Dataverse* (2019).

[109] World development indicators (2014). URL http://data.worldbank.org/indicator/NY.GDP.MKTP.CD.

[110] Marshall, M. & Gurr, T. Polity5: Political regime characteristics and transitions, 1800-2018. *Center for Systemic Peace. http://www. systemicpeace. org/inscr/p5manualv2018.pdf* (2020).

[111] Team, P. Planet application program interface: In space for life on earth (2017). URL https://api.planet.com.

[112] Team, L. Landinfo worldwide mapping llc. URL http://search.landinfo.com/.

[113] Gorelick, N. *et al.* Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* (2017). URL https://doi.org/10.1016/j.rse.2017.06.031.

[114] Jain, M. *et al.* Mapping smallholder wheat yields and sowing dates using micro-satellite data. *Remote sensing* **8**, 860 (2016).

[115] Lambert, M.-J., Blaes, X., Traoré, P. S. & Defourny, P. Estimate yield at parcel level from s2 time serie in sub-saharan smallholder farming systems. In *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, 1–7 (IEEE, 2017).

[116] Guan, K. *et al.* Mapping paddy rice area and yields over thai binh province in viet nam from modis, landsat, and alos-2/palsar-2. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **11**, 2238–2252 (2018).

[117] Karst, I. G. *et al.* Estimating yields of household fields in rural subsistence farming systems to study food security in burkina faso. *Remote Sensing* **12**, 1717 (2020).

[118] Jin, Z., Azzari, G., Burke, M., Aston, S. & Lobell, D. B. Mapping smallholder yield heterogeneity at multiple scales in eastern africa. *Remote Sensing* **9**, 931 (2017).

[119] Schulthess, U., Timsina, J., Herrera, J. & McDonald, A. Mapping field-scale yield gaps for maize: An example from bangladesh. *Field Crops Research* **143**, 151–156 (2013).

[120] Zhao, Q. *et al.* Detecting spatial variability of paddy rice yield by combining the dndc model with high resolution satellite images. *Agricultural Systems* **152**, 47–57 (2017).

[121] Lobell, D. B. *et al.* Sight for sorghums: Comparisons of satellite-and ground-based sorghum yield estimates in mali. *Remote Sensing* **12**, 100 (2020).

[122] Kim, J. H., Xie, M., Jean, N. & Ermon, S. Incorporating spatial context and fine-grained detail from satellite imagery to predict poverty (2016).

[123] Perez, A. *et al.* Poverty prediction with public landsat 7 satellite imagery and machine learning. *arXiv preprint arXiv:1711.03654* (2017).

[124] Engstrom, R., Hersh, J. & Newhouse, D. Poverty from space: using high-resolution satellite imagery for estimating economic well-being (2017).

[125] Pokhriyal, N. & Jacques, D. C. Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences* **114**, E9783–E9792 (2017).

[126] Njuguna, C. & McSharry, P. Constructing spatiotemporal poverty indices from big data. *Journal of Business Research* **70**, 318–327 (2017).

[127] Perez, A. *et al.* Semi-supervised multitask learning on multispectral satellite images using wasserstein generative adversarial networks (gans) for predicting poverty. *arXiv preprint arXiv:1902.11110* (2019).

[128] Smith, B. & Wills, S. Left in the dark? oil and rural poverty. *Journal of the Association of Environmental and Resource Economists* **5**, 865–904 (2018).

[129] Irvine, J. M., Wood, R. J. & McBee, P. Viewing society from space: Image-based sociocultural prediction models (2017).

[130] Li, G., Cai, Z., Liu, X., Liu, J. & Su, S. A comparison of machine learning approaches for identifying high-poverty counties: robust features of dmsp/ols night-time light imagery. *International journal of remote sensing* **40**, 5716–5736 (2019).

[131] Xie, M., Jean, N., Burke, M., Lobell, D. & Ermon, S. Transfer learning from deep features for remote sensing and poverty mapping. In *Thirtieth AAAI Conference on Artificial Intelligence* (2016).

[132] 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).

[133] Watmough, G. R., Atkinson, P. M., Saikia, A. & Hutton, C. W. Understanding the evidence base for poverty–environment relationships using remotely sensed satellite data: An example from assam, india. *World Development* **78**, 188–203 (2016).

[134] Mahabir, R., Agouris, P., Stefanidis, A., Croitoru, A. & Crooks, A. T. Detecting and mapping slums using open data: a case study in kenya. *International Journal of Digital Earth* **13**, 683–707 (2020).

[135] Rhinane, H., Hilali, A., Berrada, A., Hakdaoui, M. *et al.* Detecting slums from spot data in casablanca morocco using an object based approach. *Journal of Geographic Information System* **3**, 217 (2011).

[136] Hofmann, P., Strobl, J., Blaschke, T. & Kux, H. Detecting informal settlements from quickbird data in rio de janeiro using an object based approach. In *Object-based image analysis*, 531–553 (Springer, 2008).

[137] Stoler, J. *et al.* Assessing the utility of satellite imagery with differing spatial resolutions for deriving proxy measures of slum presence in accra, ghana. *GIScience & Remote Sensing* **49**, 31–52 (2012).

[138] Maiya, S. R. & Babu, S. C. Slum segmentation and change detection: A deep learning approach. *arXiv preprint arXiv:1811.07896* (2018).

[139] Duque, J. C., Patino, J. E. & Betancourt, A. Exploring the potential of machine learning for automatic slum identification from vhr imagery. *Remote Sensing* **9**, 895 (2017).

[140] Gram-Hansen, B. J. *et al.* Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 361–368 (2019).

[141] Engstrom, R. *et al.* Mapping slums using spatial features in accra, ghana. In *2015 Joint Urban Remote Sensing Event (JURSE)*, 1–4 (IEEE, 2015).

[142] Williams, T. K.-A., Wei, T. & Zhu, X. Mapping urban slum settlements using very high-resolution imagery and land boundary data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13**, 166–177 (2019).

[143] Verma, D., Jana, A. & Ramamritham, K. Transfer learning approach to map urban slums using high and medium resolution satellite imagery. *Habitat International* **88**, 101981 (2019).

[144] Leonita, G., Kuffer, M., Sliuzas, R. & Persello, C. Machine learning-based slum mapping in support of slum upgrading programs: The case of bandung city, indonesia. *Remote sensing* **10**, 1522 (2018).

[145] Graesser, J. *et al.* Image based characterization of formal and informal neighborhoods in an urban landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **5**, 1164–1176 (2012).

[146] Fallatah, A., Jones, S. & Mitchell, D. Object-based random forest classification for informal settlements identification in the middle east: Jeddah a case study. *International Journal of Remote Sensing* **41**, 4421–4445 (2020).

[147] Jochem, W. C., Bird, T. J. & Tatem, A. J. Identifying residential neighbourhood types from settlement points in a machine learning approach. *Computers, environment and urban systems* **69**, 104–113 (2018).

[148] Wurm, M., Taubenböck, H., Weigand, M. & Schmitt, A. Slum mapping in polarimetric sar data using spatial features. *Remote sensing of environment* **194**, 190–204 (2017).

[149] Kuffer, M., Pfeffer, K., Sliuzas, R. & Baud, I. Extraction of slum areas from vhr imagery using glcm variance. *IEEE Journal of selected topics in applied earth observations and remote sensing* **9**, 1830–1840 (2016).

[150] Kit, O. & Lüdeke, M. Automated detection of slum area change in hyderabad, india using multitemporal satellite imagery. *ISPRS journal of photogrammetry and remote sensing* **83**, 130–137 (2013).

[151] Hofmann, P. & Bekkarnayeva, G. Object-based change detection of informal settlements. In *2017 Joint Urban Remote Sensing Event (JURSE)*, 1–4 (IEEE, 2017).

[152] Kohli, D., Sliuzas, R. & Stein, A. Urban slum detection using texture and spatial metrics derived from satellite imagery. *Journal of spatial science* **61**, 405–426 (2016).

[153] Kohli, D., Warwadekar, P., Kerle, N., Sliuzas, R. & Stein, A. Transferability of object-oriented image analysis methods for slum identification. *Remote Sensing* **5**, 4209–4228 (2013).

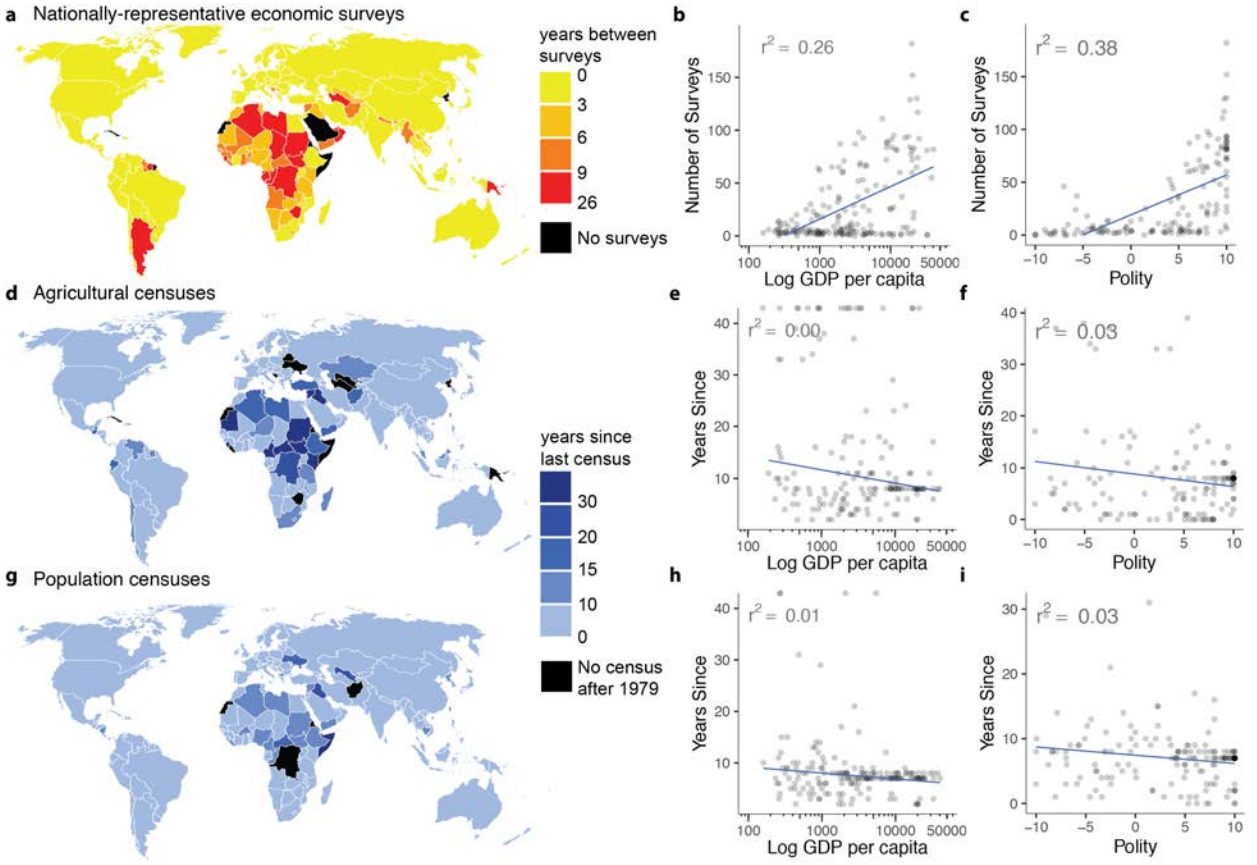

Figure 1: **Nationally-representative economic, agricultural, and population data are collected infrequently in much of the world**. **a** The average interval between nationally representative economic surveys (of Average or High quality) for the period of 1993 to 2018 from the UN World Income Inequality Database.[108] **b** Relationship between GDP per capita[109] and number of surveys in the study period. Nations with higher GDP per capita tend to have more surveys. **c** Relationship between the Polity Score of each country (+10 is fully democratic, -10 is fully autocratic)[110] and the number of surveys in the study period. **d** Years since last agricultural census. **e-f** Relationship between GDP per capita, Polity score and years since last agricultural census. **g-i** As in d-f but for population censuses.

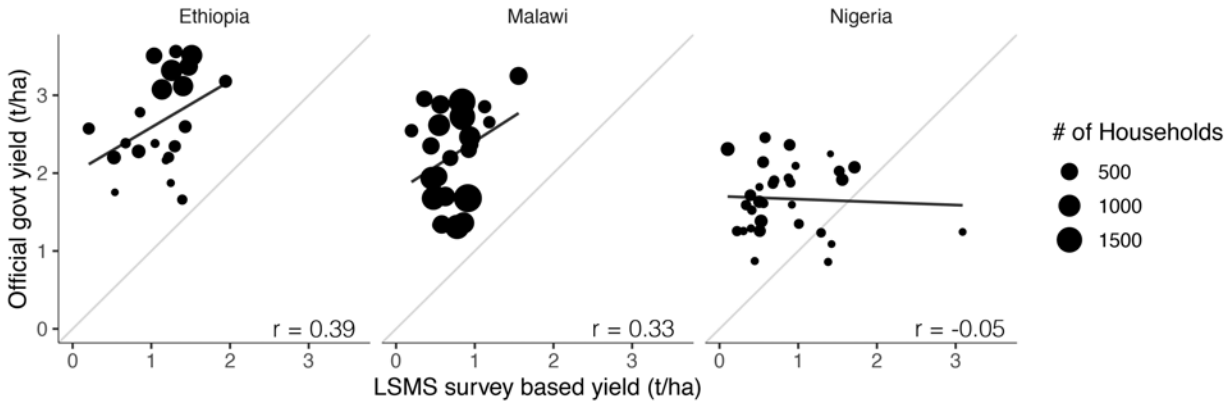

Figure 2: **Government and household-survey based data on maize productivity are not well correlated at the district level**. Using government data from eAtlas and household level yield data from LSMS-ISA surveys, maize yields are compared by averaging across all households in a given district. Data include 2011, 2013, and 2015 data in Ethiopia, 2013 data in Malawi, and 2010 and 2012 data in Nigeria. Comparison is restricted to district-years with at least 30 households. Grey line is 1:1 line, while black lines show linear fits within each country. Points are sized relative to the number of households contributing to each estimate in the LSMS data.

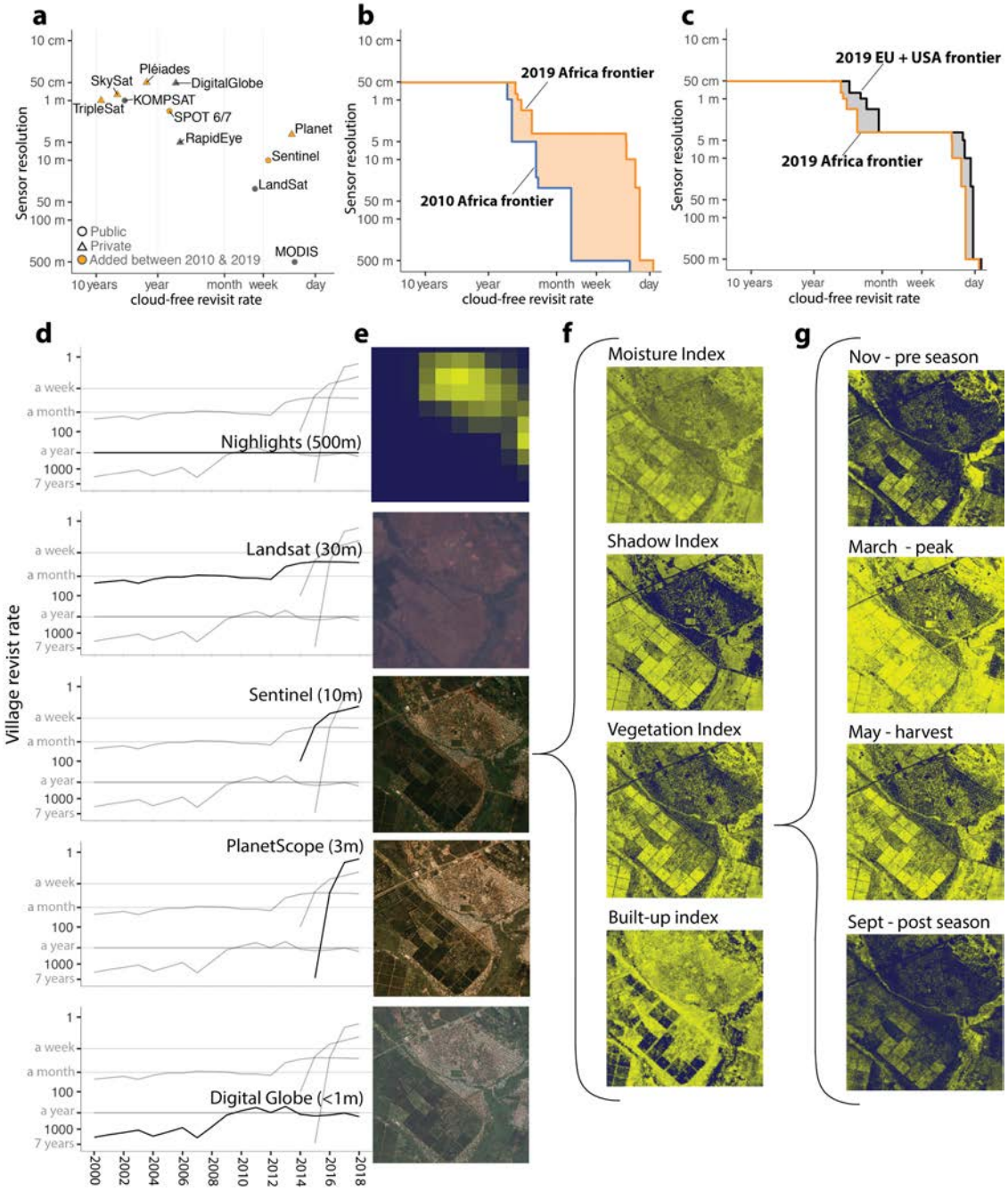

Figure 3: **Spatial resolution, temporal frequency, and spectral availability of satellite imagery have increased substantially since 2000**. **a** Average revisit rate and sensor resolution of cloud-free optical imagery in 2019, averaged across 100 populated African locations. **b** Blue line ("frontier") shows overall revisit rate across all available sensors at a given spatial resolution in 2010 for same 100 locations (e.g. at 1m, line denotes average cloud-free revisit rate using all sensors $<=$1m); orange line shows same for 2019. Orange area denotes the new combinations of temporal and spatial resolution available by 2019, which expanded greatly at resolutions $>$1m **c** Average 2019 coverage in Africa (orange line) vs 100 locations in US/EU (grey line; locations randomly sampled proportional to population). Grey shaded area depicts inequalities in coverage between wealthy and developing regions in 2019, which are larger for imagery $<$3m/px. **d** Calculated revisit periods for several satellites over 500 randomly selected survey locations in Africa since 2000. Nightlights is set to a one year revisit rate given the stable yearly product. **e** Example imagery corresponding to each sensor in a single location in central Zambia. Images are real color except for NL. **f** Indices generated from various bands can convey different information, as depicted here using Sentinel 2 data (yellow colors indicate higher values of the index). **g** Frequent revisit rates of new public sensors capture temporal variation in human activity, including rapid changes throughout the main agricultural season shown here.

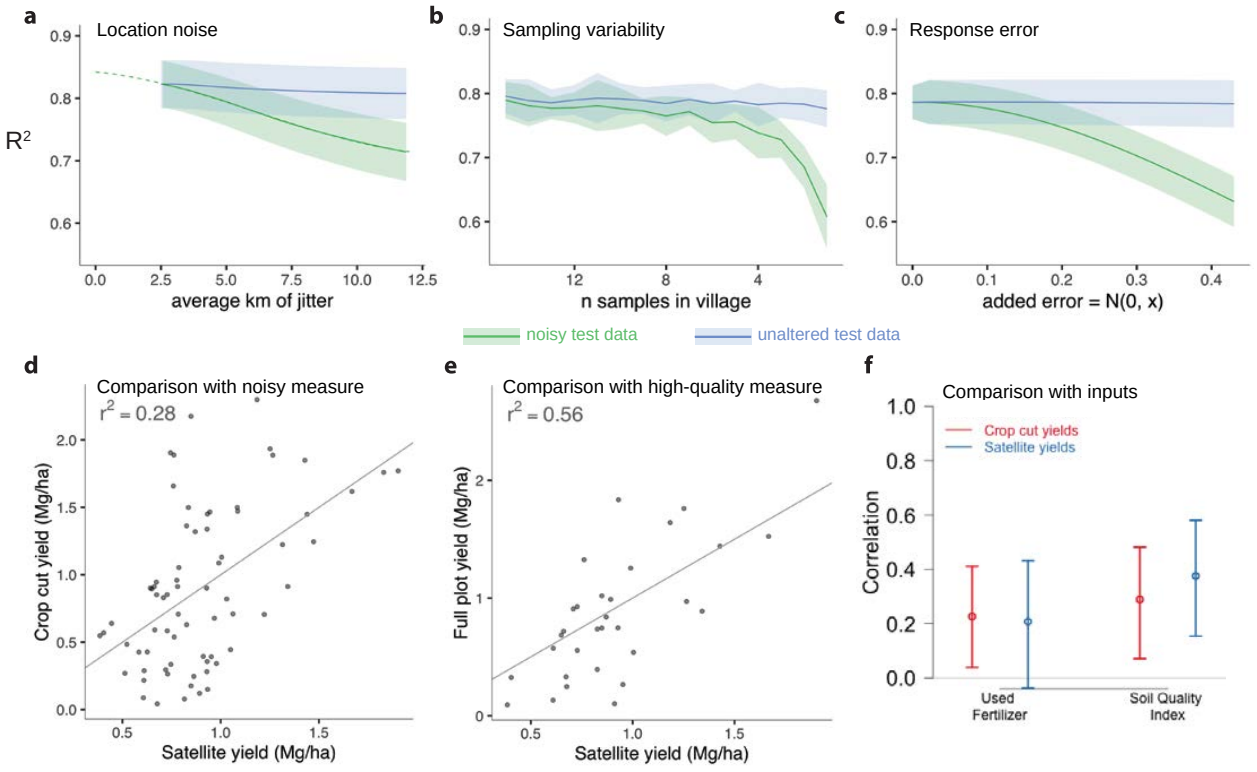

Figure 4: **The role of noise in model performance and evaluation**. **a-c** Performance of wealth prediction model as noise is added to train and test data. Model trained to predict asset wealth from nightlights imagery across 4000 African villages, using the dataset from ref.[30] Performance is evaluated as three different types of noise are added to training data: **a** random noise in village geo-coordinates (starting from 2.5km, the actual noise in the survey data), **b** noise from constructing village-level wealth estimates from decreasing numbers of households within the village to represent sampling variability, and **c** random noise added to village-level wealth estimates, representing random response error from respondents. Green lines show performance evaluated on test data where similar noise has been added, blue lines show performance on test data where noise has not been added. Shaded areas indicate confidence intervals across 200 runs at a given level of added noise. As all types of training noise increase, model performance degrades when evaluated against similarly noisy test data but does not degrade when evaluated against unaltered test data. **d-f** Example from a study of maize yields in Uganda[54] in which both ground-based and satellite-based measurements can have noise, and multiple approaches can help adjudicate which is noisier. **d** Imperfect correlation between ground- and satellite-based yield measure does not reveal source of noise. **e** Comparison of satellite measure with available gold-standard ground measure from full plot harvest shows higher correlation, indicating ground measure in (d) responsible for at least some of the noise. **f** Comparison of satellite measure and ground measure with independent third measures expected to correlate with yields (here, fertilizer use and soil quality) suggests that the two yield measures in (d) are roughly equally noisy.

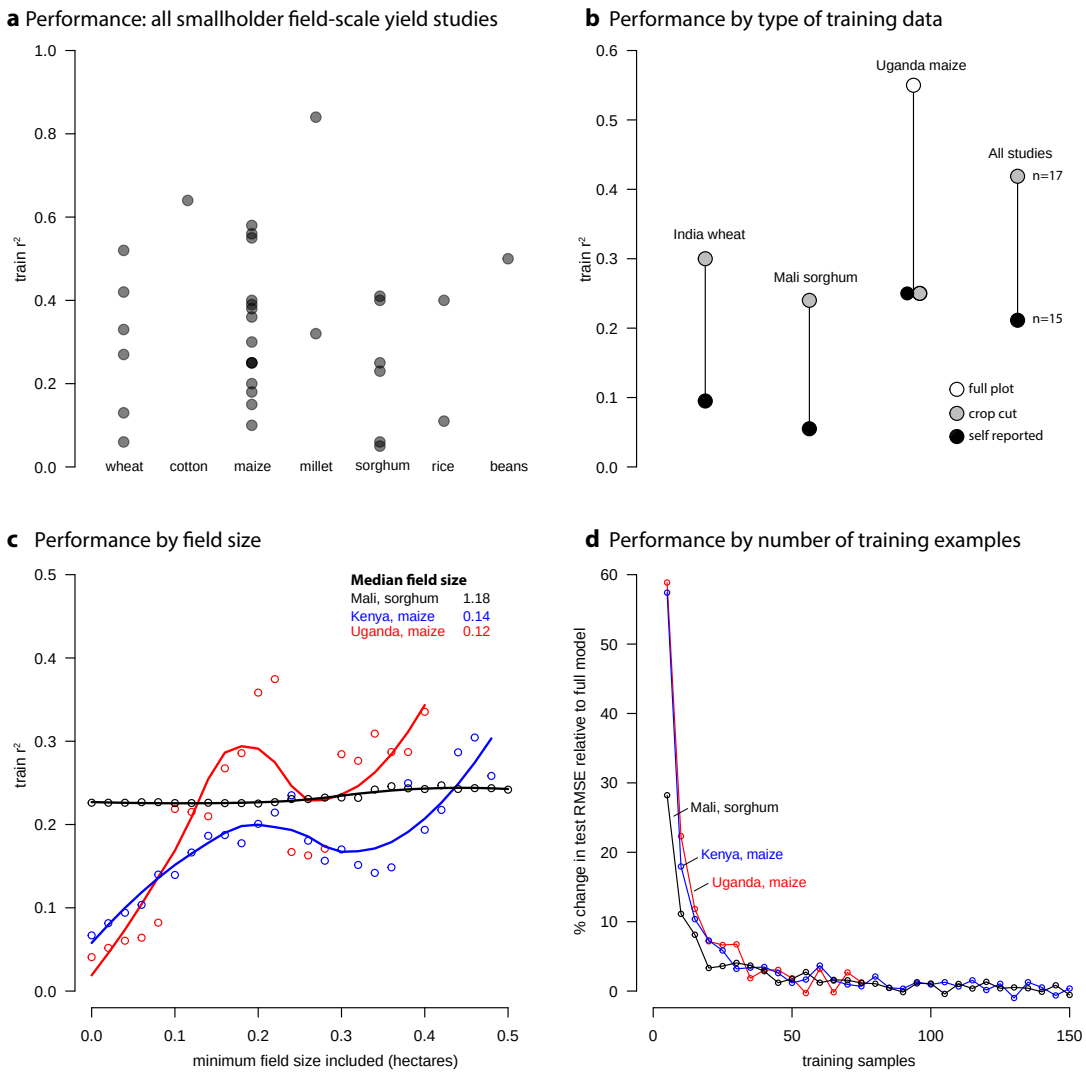

Figure 5: **Performance of satellite-based approaches to measuring smallholder yield at field scale**. **a** Performance across all known published studies where coefficient of determination (r^2) was reported (32 estimates across 11 studies); r^2 estimates are "in-sample", i.e. for data on which model was trained. **b** Difference in performance for models trained and evaluated on crop-cut, self-reported, or full-plot harvest data suggest that more objective crop measures improve performance. First three estimates are for studies that compared at least two types of ground data in the same setting. "All studies" estimates pool across estimates in (a). **c** Performance generally increases when sample is restricted to larger fields, particularly in East African settings where field sizes are very small. **d** Performance on test data improves rapidly with additional training examples up to $\sim$30 data points, and then improves more gradually thereafter. Performance measured as average root mean squared error between predicted and observed yields in the test set, averaged over 100 different random subsets of training samples at each size of the training set.

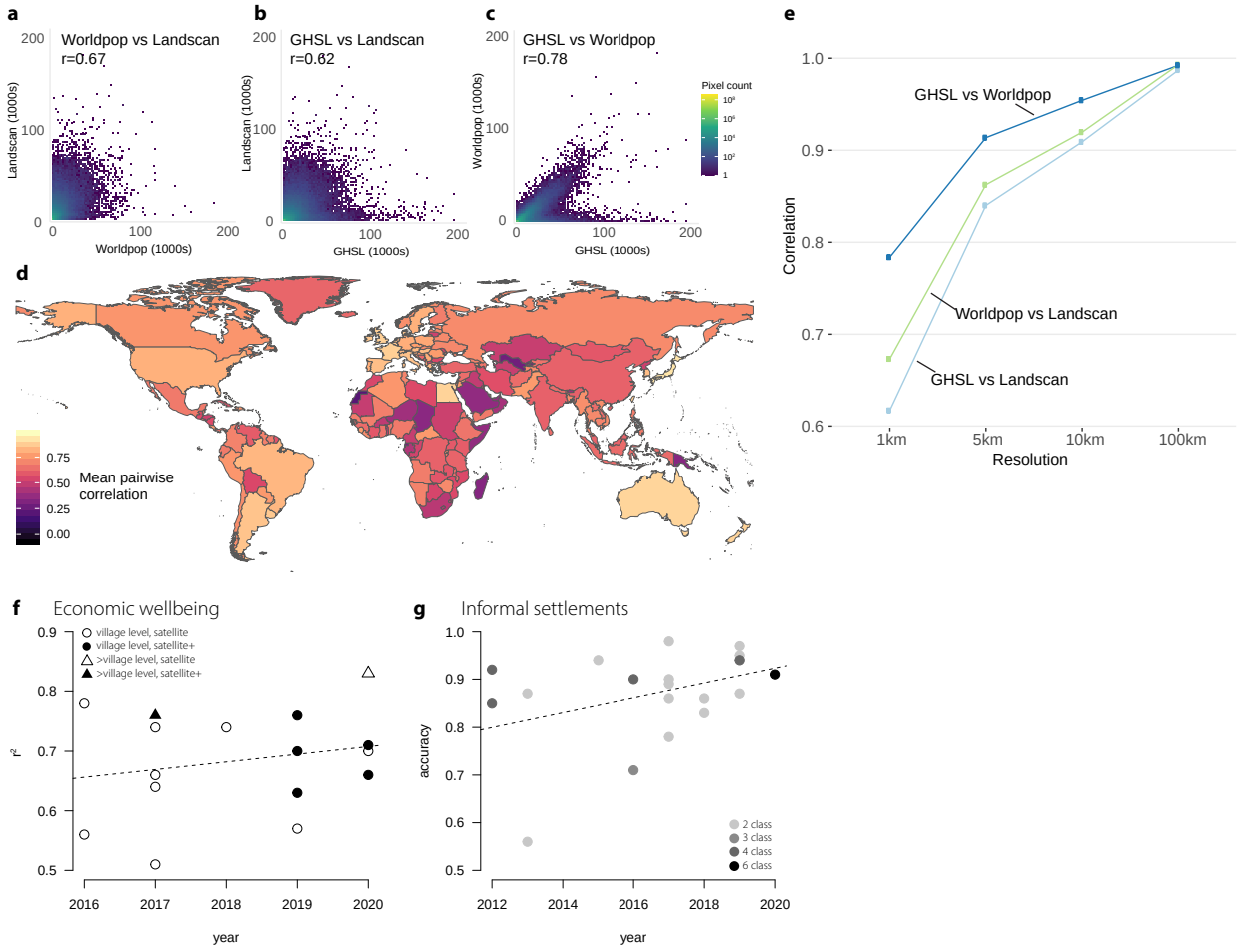

Figure 6: **Performance of satellite-based approaches to measuring population, wealth, and informal settlements**. **a-c** Comparison of three different satellite-informed global population datasets (Landscan, WorldPop, and GHSL population) datasets at 1km resolution globally (colors correspond to scale at right). **d** Average pairwise correlation within each country at 1km resolution. Comparisons show modest correlation between datasets at global scale and often poor correlation in many developing countries. **e** Correlations across datasets improve when data are spatially aggregated. All comparisons are made for pixels that were not missing and not zero across all three datasets. **f** Performance in predicting asset wealth in various developing countries from satellite data (16 estimates from 12 papers), as measured by coefficient of determination on test data. Filled markers are estimates that combine satellite information with other data (cell phone data, social media data, or Wikipedia). Circles indicate estimates at the village level, triangles are estimates at more aggregate spatial scale (sub-district or district). **g** Performance in predicting the location of informal settlements from imagery (20 estimates from 17 papers

.

Supplementary Information

Collecting Satellite Revisit Data

Construction of Figure 3 involved acquiring data from several sources. First we use Gridded Population of the World (GPW) data raster data to create a population weighted sample of 100 locations in Africa, as well as 100 locations across the EU and the USA. These 200 locations are then buffered with an approximately 10 meter radius and used to query for satellite imagery for the years 2010 and 2019.

Planet products (SkySat, PlanetScope and RapidEye) were downloaded from the Planet API.[111] Footprints for other private satellites were downloaded from LandInfo,[112] while footprints from public satellites (Landsat, Sentinel, MODIS) were downloaded using Google Earth Engine.[113]

We attempted to maintain consistency in filtering across data sources, but filtering works slightly differently in each system. LandInfo (though it lacks in depth documentation confirming this) appears to filter solely over the area of interest (AOI) rather than over the entire footprint. Public data was processed to match this, using only the cloud cover percentage over the buffered polygon, but Planet is filtered at the footprint level. All image footprints were filtered to be <30% cloud cover and off-nadir <|20|.

Sensors in Figure 3a are grouped slightly to make the figure easier to process visually. Landsat 7 and 8 are combined; WorldView-1 through 4, GeoEye-1, QuickBird-2, and IKONOS are grouped as “DigitalGlobe"; KOMPSAT-3, KOMPSAT-3A, and KOMPSAT-2 are grouped; SPOT-4 and 5 are grouped as well as SPOT-6 and 7. As sensor resolution varies within these groups, we use the mode of the resolutions in the group to represent the group as a whole. This does compress the range of resolutions significantly, for example “DigitalGlobe" is recorded as a resolution of 51cm, where the true resolutions range from 31cm to 91cm. “KOMPSAT" ranges from 70cm to 1m, “SkySat" ranges from 70cm to 1m, “SPOT 4/5" ranges from 2.5m to 10m.

To calculate the average revisit rate, we sum up the total number of images collected in each group and calculate (number of locations*365)/number of images. For the frontier, we calculate the revisit rate by summing the total number of images collected for all satellites with resolution less than or equal to the resolution of interest and run the same calculation as above. As this is an average of time between images, a number below 1 does not necessarily indicate that there is a cloudless picture on every day.

Table S1: Performance of studies using satellites to predict smallholder yields at the plot level. Year=study year, res=sensor resolution, n=number of observations, r2=squared correlation, data=training data (CC=crop cuts, SR=self reported, FP=full plot). In the sensor column, “Skysat (c)" refers to Skysat data that has been coarsened to lower resolution.

	study	year	location	crop	sensor	res	n	r2	data
1	Jain et al 2016[114]	2014	India	wheat	Skysat	2	50	0.27	CC
2	Jain et al 2016[114]	2015	India	wheat	Skysat	2	37	0.33	CC
3	Jain et al 2016[114]	2014	India	wheat	Skysat	2	52	0.13	SR
4	Jain et al 2016[114]	2015	India	wheat	Skysat	2	29	0.06	SR
5	Jain et al 2016[114]	2014	India	wheat	Landsat	30	50	0.52	CC
6	Jain et al 2016[114]	2015	India	wheat	Landsat	30	37	0.42	CC
7	Lambert et al 2017[115]	2016	Mali	cotton	S2	10	9	0.64	CC
8	Lambert et al 2017[115]	2016	Mali	maize	S2	10	9	0.58	CC
9	Lambert et al 2017[115]	2016	Mali	millet	S2	10	8	0.84	CC
10	Lambert et al 2017[115]	2016	Mali	sorghum	S2	10	9	0.41	CC
11	Guan et al 2017[116]	2015	Vietnam	rice	Landsat	30	71	0.40	CC
12	Karst et al 2020[117]	2018	Burkina Faso	beans	S2	10	31	0.50	CC
13	Karst et al 2020[117]	2018	Burkina Faso	maize	S2	10	31	0.40	CC
14	Karst et al 2020[117]	2018	Burkina Faso	sorghum	S2	10	57	0.40	CC
15	Karst et al 2020[117]	2018	Burkina Faso	millet	S2	10	45	0.32	CC
16	Jin et al 2017[118]	2016	Kenya	maize	S2	10	41	0.36	CC
17	Schulthess et al 2013[119]	2010	Bangladesh	maize	rapideye	5	30	0.56	SR
18	Zhao et al 2017[120]	2009	China	rice	formosat-2	8	22	0.11	SR
19	Lobell et al 2020[121]	2017	Mali	sorghum	S2	10	575	0.23	CC
20	Lobell et al 2020[121]	2017	Mali	sorghum	Planet	3	575	0.25	CC
21	Lobell et al 2020[121]	2017	Mali	sorghum	S2	10	575	0.05	SR
22	Lobell et al 2020[121]	2017	Mali	sorghum	Planet	3	575	0.06	SR
23	Burke and Lobell 2017[16]	2014	Kenya	maize	Skysat	1	72	0.39	SR
24	Burke and Lobell 2017[16]	2014	Kenya	maize	Skysat (c)	5	72	0.38	SR
25	Burke and Lobell 2017[16]	2014	Kenya	maize	Skysat (c)	10	72	0.30	SR
26	Burke and Lobell 2017[16]	2014	Kenya	maize	Skysat (c)	30	72	0.25	SR
27	Burke and Lobell 2017[16]	2014	Kenya	maize	Skysat	1	386	0.20	SR
28	Burke and Lobell 2017[16]	2014	Kenya	maize	Skysat (c)	5	386	0.18	SR
29	Burke and Lobell 2017[16]	2014	Kenya	maize	Skysat (c)	10	386	0.15	SR
30	Burke and Lobell 2017[16]	2014	Kenya	maize	Skysat (c)	30	386	0.10	SR
31	Lobell et al 2019[54]	2016	Uganda	maize	S2	10	252	0.55	FP
32	Lobell et al 2019[54]	2016	Uganda	maize	S2	10	252	0.25	SR
33	Lobell et al 2019[54]	2016	Uganda	maize	S2	10	252	0.25	CC

Table S2: Performance of efforts to predict economic wellbeing with imagery and ML. “Samples" reports the number of training examples when available. “Geo" reports the geographic level at which models were evaluated, with (e.g.) "geo2" equal to county or district.

	Year	Target	Metric	Result	Samples	Geo	Location
[30]	2020	Asset Wealth Index	r2	0.83	19,669 clusters	geo2	23 African countries
[122]	2016	Asset Wealth Index	r2	0.78		cluster	Nigeria
[29]	2017	Wealth Index	r2	0.76		voroni polygons	Bangladesh
[27]	2016	Asset Wealth Index	r2	0.75	1,411 clusters	cluster	Rwanda
[123]	2017	Asset Wealth Index	r2	0.66		cluster	Africa
[124]	2017	Income <= National 10th percentile	r2	0.61	1,291 villages	geo4	Sri Lanka
[57]	2020	Consumption	r2	0.54	320 clusters	cluster	Uganda
[125]	2017	Multi-Dimensional Poverty Index	cor	0.91	552 communes	geo4	Senegal
[126]	2016	Multi-Dimensional Poverty Index	cor	0.88	416 sectors	geo3	Rwanda
[127]	2019	Asset Wealth Index	cor	0.57	4,839 clusters	cluster	Africa
[128]	2017	If below Comparative Wealth Index poverty line or not	accuracy	0.83	eval on 636,448 hholds	cluster	36 countries
[129]	2017	Reported living condition good/neutral/bad	accuracy	0.83		cluster	Botswana, Kenya, Zimbabwe
[130]	2019	If county is "non-poverty" or not	accuracy	0.82	192 counties	geo2	China
[131]	2016	Above or below poverty line	accuracy	0.72	643 clusters	cluster	Uganda
[132]	2018	Bottom 40%, middle 40%, top 20% classification	accuracy	0.62	330 hholds	hhold	Kenya
[133]	2016	Welfare Index quintiles	accuracy	0.36	14,000 clusters	cluster	India

Table S3: Performance of efforts to predict the location of informal settlements (slums) with imagery and ML

	Year	Target	Metric	Result	Samples	Location
[134]	2018	MajiData spatial extent of slums	recall	0.95		Kenya
[135]	2011	Slum delineations	recall	0.85	70km^2 classified	Morocco
[136]	2008	Manual slum delineations	recall	0.68		Brazil
[137]	2012	Slum index	r2	0.4	eval on 1,724 EAs	Ghana
[138]	2018	Slum Delineations	IoU	0.9		India
[139]	2017	Slum delineations	accuracy	0.98	12,398 100m cells	Colombia
[140]	2019	Annotated ground truth points for slums	accuracy	0.97		India
[141]	2015	Accra Metropolitan Assembly slum dichotomy map	accuracy	0.94	3,000 samples	Ghana
[142]	2019	Point locations of squatter settlements	accuracy	0.94		Jamaica
[143]	2019	Slum delineations	accuracy	0.94		India
[144]	2018	Slum delineations	accuracy	0.94		Indonesia
[145]	2012	?	accuracy	0.92	12,000 points	Afghanistan
[146]	2020	Slum delineations	accuracy	0.91		Saudi Arabia
[36]	2017	Slum delineations	accuracy	0.9	3,000 samples	Tanzania
[35]	2017	Slum delineations	accuracy	0.9	3,060 samples	Tanzania
[147]	2018	Regular/irregular settlement 1km grid	accuracy	0.9		Afghanistan
[148]	2017	Slum areas from visual image interpretation	accuracy	0.89	1,159,662 pixels	India
[149]	2016	Municipality provided slum location data	accuracy	0.88	80 points	India
[150]	2013	Rule based classification, using 7 ground truth locations	accuracy	0.87	eval on 7 points	India
[151]	2017	Manual classification map	accuracy	0.78	eval on region	South Africa
[152]	2016	Slum delineations	accuracy	0.71	eval on 250 pts	India
[153]	2013	Manual slum delineations	accuracy	0.56	eval on city	India