REVIEW SUMMARY

SUSTAINABILITY

Using satellite imagery to understand and promote sustainable development

Marshall Burke*, Anne Driscoll, David B. Lobell, Stefano Ermon

BACKGROUND: Accurate and comprehensive measurements of a range of sustainable development outcomes are fundamental inputs into both research and policy. For instance, good measures are needed to monitor progress toward sustainability goals and evaluate interventions designed to improve development outcomes. Traditional approaches to measurement of many key outcomes rely on household surveys that are conducted infrequently in many parts of the world and are often of low accuracy. The paucity of ground data stands in contrast to the rapidly growing abundance and quality of satellite imagery. Multiple public and private sensors launched in recent years provide temporal, spatial, and spectral information on changes happening on Earth's surface.

Here we review a rapidly growing scientific literature that seeks to use this satellite imagery to measure and understand various outcomes related to sustainable development. 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 and enable new insights. Our focus is on settings and applications where humans themselves, or what they produce, are the outcome of interest and on where these outcomes are being measured using satellite imagery.

ADVANCES: We describe and synthesize the variety of approaches that have been used to extract information from satellite imagery, with particular attention given to recent machine learning–based approaches and settings in which training data are limited or noisy. We then quantitatively assess predictive performance of these approaches in the domains of smallholder agriculture, economic livelihoods, population, and informal settlements. We show that satellite-based performance in predicting these outcomes is reasonably strong and improving. Performance improvements have come through a combination of more numerous and accurate training data, more abundant and higher-quality imagery, and creative application of advances in computer vision to satellite inputs and sustainability outcomes. Further, our 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. We describe multiple methods through which the true performance of satellite-based approaches can be better understood.

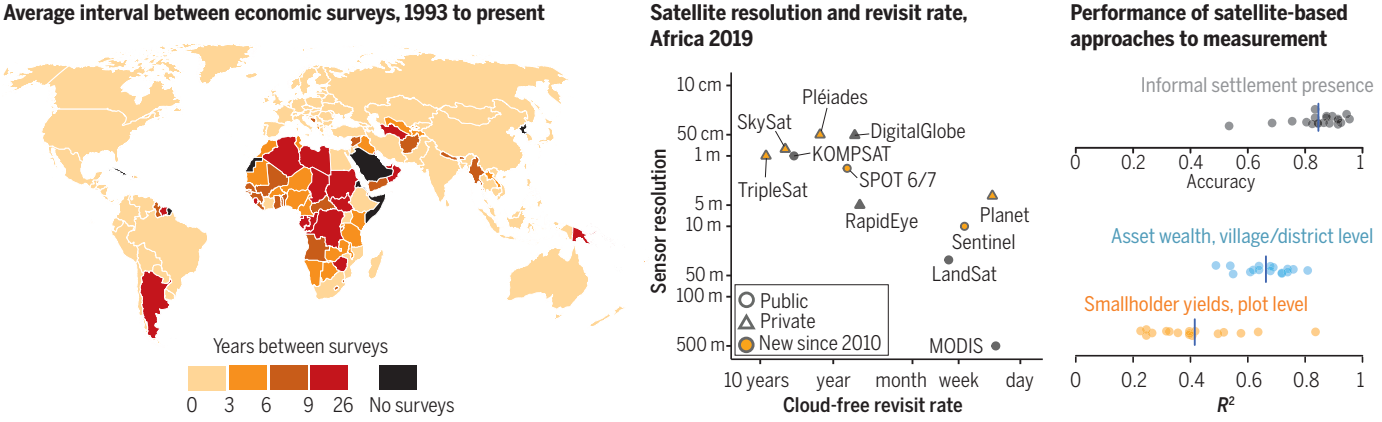
Integration of satellite-based sustainability measurements into research has been broad, and we describe applications in agriculture, fisheries, health, and economics. Documented uses of these measurements in public-sector decision-making are rarer, which we attribute in part to the novelty of the approaches, their lack of interpretability, and the potential benefits to some policy-makers of not having certain outcomes be measured.

OUTLOOK: The largest constraint to satellite-based model performance is now training data rather than imagery. While imagery has become abundant, the scarcity and frequent unreliability of ground data make both training and validation of satellite-based models difficult. Expanding the quantity and quality of such data will quickly accelerate progress in this field. Other opportunities for advancement include improvements in model interpretability, fusion of satellites with other nontraditional data that provide complementary information, and more-rigorous evaluation of satellite-based approaches (relative to available alternatives) in the context of specific use cases.

Nevertheless, despite the current and future promise of satellite-based approaches, we argue that these approaches will amplify rather than replace existing ground-based data collection efforts in most settings. 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. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: mburke@stanford.edu
Cite this article as M. Burke *et al.*, *Science* **371**, eabe8628 (2021). DOI: 10.1126/science.abe8628

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abe8628



Increasing collection of satellite imagery can help measure livelihood outcomes in areas where ground data are sparse. (Left) Interval between nationally representative economic surveys over the past three decades shows long lags in many developing countries. (Middle) Recently added public and private satellites have broken the traditional trade-off between temporal and spatial resolution. (Right) Performance in measuring the presence of informal settlements, crop yields on smallholder agricultural plots, and village-level asset wealth. R^2, coefficient of determination.

REVIEW

SUSTAINABILITY

Using satellite imagery to understand and promote sustainable development

Marshall Burke[1,2,3]*, Anne Driscoll[2], David B. Lobell[1,2], Stefano Ermon[4]

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 improving 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 model performance. We quantify recent model performance across multiple sustainable development domains, discuss research and policy applications, explore constraints to future progress, and highlight research directions for the field.

Humans have long sought to image their habitat from above the ground. Socrates purportedly stated in 500 BCE 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-19th century CE, with earth observation data collected by strapping cameras to balloons, kites, airplanes, 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 nonmilitary earth observation satellites in orbit, 75% of which were launched within the past five years (*3*). These satellites are now capturing imagery of Earth with 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 United Nation's Sustainable Development Goals (*4*). 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 nontraditional, unstructured data (e.g., data from social media or cell phones) to measure human-related outcomes unless these data are combined with imagery. Our review complements existing sector-specific reviews, including the use of remote sensing in agriculture (*5*, *6*), in economic applications (*7*), and in the detection of informal settlements (*8*), drawing common lessons across these and other domains.

We make 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. 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 ground data make both training and validation of satellite-based models difficult. Third, despite the growing power of satellite-based approaches, these approaches will likely amplify rather than fully replace existing ground-based data collection efforts, given the necessity of training data and the likelihood that many outcomes of interest will likely never be accurately estimated with satellites.

Finally, in the sustainable development domains on which we focus, there remain few documented cases where satellites have been used in public-sector decision-making processes—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.

The availability and reliability of data

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

Household- or field-level surveys remain the main data collection tool for key development-related 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. But their implementation and use also face a number of important challenges. First, nationally representative surveys are expensive and time-consuming to conduct. Conducting a Demographic and Health Survey (DHS) or Living Standards Measurement Study (LSMS) in one country for one year typically costs $1.5 million to $2 million USD (*9*), 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 US dollars in a typical African country (*10*).

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 (Fig. 1A), as compared with subannual frequency in most wealthy countries. Survey frequency is on average substantially lower in less wealthy countries (Fig. 1B), meaning that data on livelihood outcomes are often lacking where they are arguably 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 (*11*). The frequency of agricultural and population censuses also varies widely around the world (Fig. 1, D and G).

[1]Department of Earth System Science, Stanford University, Stanford, CA, USA. [2]Center on Food Security and the Environment, Stanford University, Stanford, CA, USA. [3]National Bureau of Economic Research, Cambridge, MA, USA. [4]Department of Computer Science, Stanford University, Stanford, CA, USA.
*Corresponding author. Email: mburke@stanford.edu

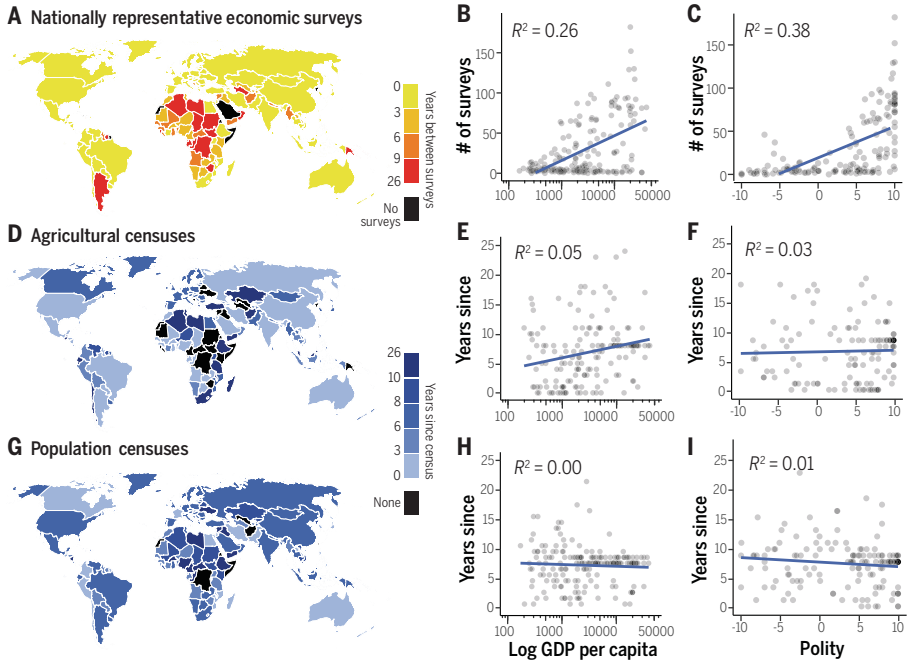


Fig. 1. Nationally representative economic, agricultural, and population data are collected infrequently in much of the world. (**A**) The average interval between nationally representative economic surveys (of average or high quality) for the period 1993–2018 from the UN World Income Inequality Database (*109*). (**B**) Relationship between gross domestic product (GDP) per capita (*110*) 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) (*111*) and the number of surveys in the study period. (**D**) Years since last agricultural census, using data covering 1993–2018. (**E** and **F**) Relationship between GDP per capita, Polity score, and years since last agricultural census. (**G** to **I**) As in (D) to (F), but for population censuses.

For instance, 24% of the world's countries (49 out of 206) have gone more than 15 years since their last agricultural census, and 6% (13 out of 206) have gone more than 15 years since their last population census.

A second challenge is that survey samples are typically only representative at the national or (sometimes) regional level, meaning that 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 example, targeting an antipoverty program or studying its impact.

Third, underlying data 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 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 noise has important implications for how satellite-based models trained on these data are validated and interpreted.

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 are inferred from responses to surveys. These responses can introduce large amounts of both random and systematic measurement error. 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 (*12*). In agriculture, 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 (*13*). Current practices are likely to have a bias toward overestimation, further weakening the quality of food security assessments (*13*, *14*).

An additional key source of noise comes from sampling variability. 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 Fig. 2, we compare average maize yields at the first administrative level (e.g., province or state) as obtained from household surveys covered by the LSMS–Integrated Surveys on Agriculture (ISA) program versus by official government ministry estimates in three African countries. This comparison reveals both a systematic bias toward higher yields in official government data than in household responses and a relatively low correlation between the two measures, with the highest observed correlation coefficient r of 0.39 for Ethiopia.

A third common source of error 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 2 km of random jitter added in urban areas and 5 km in rural areas. Below we explore the implications of these three sources of error for model development and evaluation.

Availability of satellite imagery changing rapidly

Information from satellite imagery has been used to a limited extent in both agricultural and socioeconomic applications for decades (*15*, *16*). 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 and a corresponding use of this imagery in applications.

To quantify this increase in imagery and understand how it varies across developing and developed countries, we randomly sampled 100 locations in Africa and 100 additional locations across the US and EU (sampling proportional to population) and queried the availability of cloud-free imagery (defined as <30% cloud cover) at each location in 2010

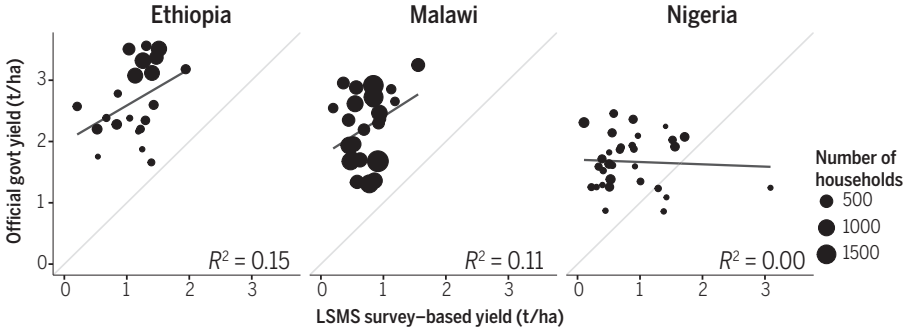


Fig. 2. Government and household survey–based data on maize productivity are not well correlated at the district level. Using government data from eAtlas and household-level yield data from LSMS-ISA surveys, maize yields (metric tons per hectare) are compared by averaging across all households in a given 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. Gray 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.

and 2019 for all available optical sensors, using multiple online query tools (*17*). We calculated region- and year-specific average revisit rates for each sensor and constructed an imagery-resolution "frontier," defined as the overall revisit rate across sensors at or below a given spatial resolution.

We found that the addition of many new sensors has lessened the traditional trade-off between temporal and spatial resolution (Fig. 3A), particularly at resolutions ≥3 m. Although the revisit rate of very-high-resolution (<1 m) sensors over Africa has seen only slight improvement over the past decade (Fig. 3B), and very-high-resolution revisit rates remain lower in Africa than in both the US and EU (Fig. 3C), revisit rates for high-resolution (1 to 5 m) and moderate- to low-resolution sensors have increased drastically and are globally equitable.

We sampled and visualized additional images and sensors across populated African locations (Fig. 3). Various types of human activity are readily visible even with moderate-resolution sensors (5 to 30 m), 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).

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 crop yield estimates for an area). Such models are often simple, such as linear regression models that relate satellite-derived vegetation indices to crop yields (*18*) or that relate nighttime lights (henceforth, "nightlights") to economic outcomes (*19*). 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. 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 handcrafted 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, 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 handcrafted features to the outputs of interest [see (*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 (e.g., minimum, median, or maximum), or histograms of binned values as inputs to a regression model. For example, Henderson *et al.* (*19*) showed how average growth in nightlights over a country was a strong predictor of economic growth at the country level in a linear regression, and our previous work (*18*) showed how vegetation indices derived from daytime high-resolution imagery were strong predictors of smallholder maize yields in a linear regression.

Models that use spatial structure in the imagery

In computer vision, spatial context can often greatly improve prediction accuracy for image 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 handcrafted features and simple aggregation strategies. Deep networks with residual connections such as DenseNET or ResNet (*21*) are often used. In this case, features are automatically learned from the data rather than handcrafted. This is currently the leading approach in most computer vision applications. Use of this approach with satellite images in sustainable development applications has proliferated in recent years, including in the measurement of population (*22*–*24*), economic livelihoods (*25*–*28*), infrastructure quality (*29*, *30*), land use (*31*, *32*), informal settlements (*33*, *34*), fishing activity (*35*, *36*), and many others. In one example, a team hand-annotated thousands of medium-resolution daytime images with the location of foreign fishing vessels and then trained a CNN to predict the presence of those vessels; predictions were then further hand-validated using high-resolution imagery (*36*).

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., ambiguity due to cloud cover) and 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 long short-term memory networks (LSTMs) (*37*), convolutional LSTMs (*38*), and three-dimensional (3D) CNNs, where images are fed in sequence into the model before it makes a prediction. These models have been successfully used for crop classification (*39*–*41*), crop yield prediction (*42*, *43*), predicting landslide susceptibility (*44*), and assessing building damage after disasters (*45*, *46*), among many other applications. For example, You *et al.* (*42*) assemble near-daily coarse-resolution multispectral images across the US Midwest, convert each band in each image to a histogram of reflectance values, and then train both a 3D CNN and an LSTM to predict county-level soybean yields from

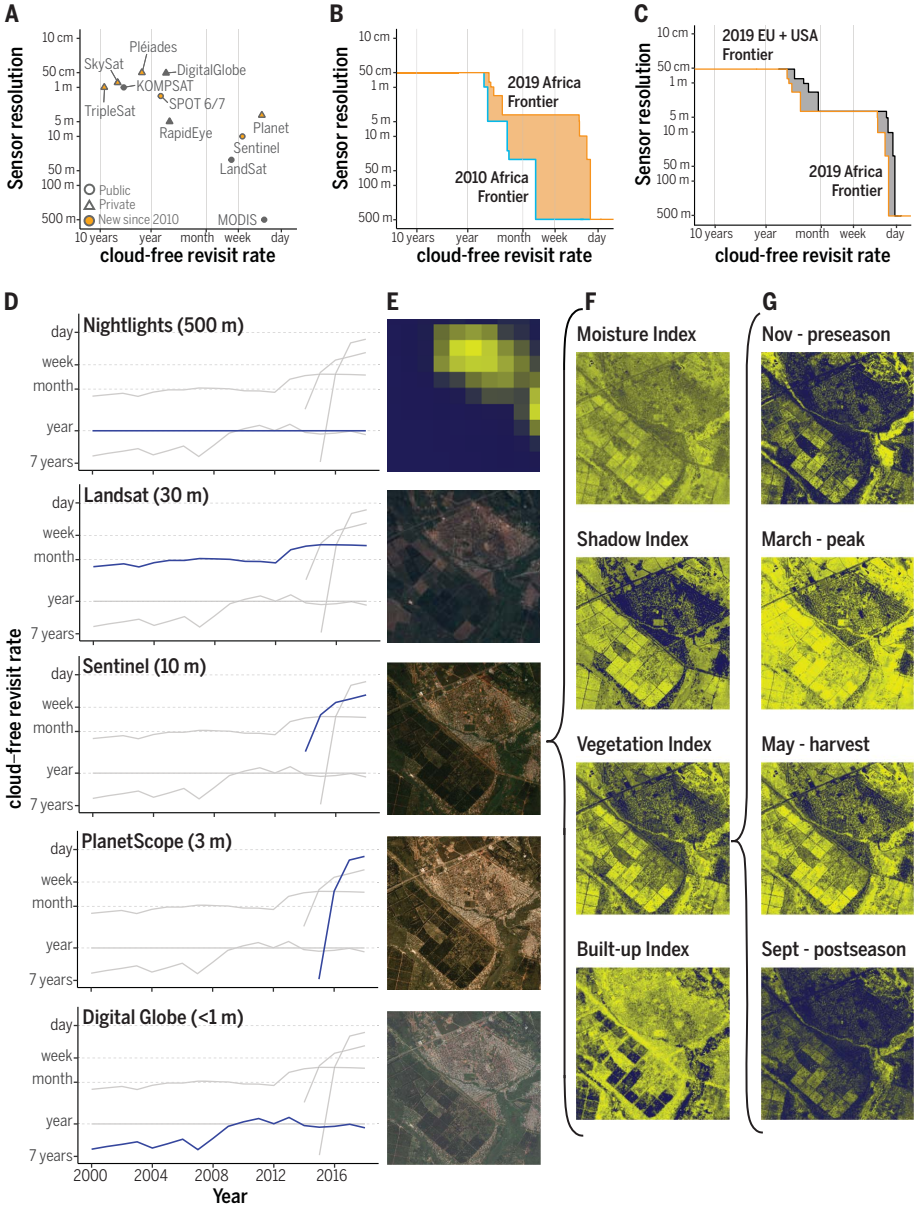


Fig. 3. Spatial resolution, temporal frequency, and spectral availability of satellite imagery have increased substantially since 2000. (**A**) Average revisit rate and sensor resolution of cloud-free optical imagery in 2019, averaged across 100 populated African locations (randomly sampled, proportional to population). (**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 1 m, the line denotes average cloud-free revisit rate using all sensors ≤1 m); 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 >1 m. (**C**) Average 2019 coverage in Africa (orange line) versus 100 locations in US or EU (gray line; locations randomly sampled, proportional to population). Gray shaded area depicts inequalities in coverage between US or EU and Africa in 2019, which are larger for imagery <3 m per pixel. (**D**) Calculated revisit periods for several satellites over 500 randomly selected survey locations in Africa since 2000. Nightlights revisit rate is set to 1 year 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 nightlights. (**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.

these histograms, outperforming previous models.

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 that are then concatenated before a final prediction layer. One example of this approach is a model that combined both daytime and nighttime satellite imagery to predict village-level asset wealth in Africa (*28*); separate CNNs were trained to predict wealth using the two types of imagery, and then the final layers of each model were concatenated and used as predictors in a final ridge regression. Additional examples include models that combine imagery with data from weather sensors (*47*), cell phones (*27*), Wikipedia (*48*), social media (*49*), street-level imagery (*50*), or Open Street Map (*51*) to predict development-related outcomes.

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 (*52*), 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. 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 can be used to estimate relationships between crop yields and physical parameters (e.g., leaf area index, canopy nitrogen) that have expected relationships with vegetation indices; model parameters can then be combined with observed vegetation indices from satellite imagery to predict yields. This approach requires no ground data for training and has been shown to perform as well as or better than approaches that calibrate directly to limited field data (*18*, *53*).

Transfer learning

A second approach, transfer learning, first trains a model 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 (*54*) and WikiSatNet (*55*) for satellite images]. The model is then "fine-tuned" on the target task of interest. For example, Jean *et al.* (*25*) first trained a neural network 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 a very small (<500) number of observations of economic livelihoods in Africa 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 well-being in Uganda with high accuracy (*56*).

Transfer learning can also be done spatially, with models trained in a region with plentiful data and then fine-tuned to a target geography where labels are sparse. For example, a model trained to predict infrastructure quality in Africa was fine-tuned to a specific country using only a small amount of labeled data (*30*). 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 to take advantage of large amounts satellite imagery for which labels are not available. Pretraining on unlabeled data has shown great progress in computer vision (*57*–*59*), narrowing the gap with fully supervised methods. For instance, two related recent approaches train CNNs to use spatial similarity between small patches of input images to derive representations of entire images without any labeled training data. Representations learned this way perform well on a range of tasks, such as crop-type classification and prediction of wealth and population (*60*, *61*). Semi-supervised learning strategies attempt to improve model performance by additionally leveraging a small amount of labeled data. This idea improved performance in predicting economic well-being from satellite imagery (*62*).

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 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 predictive features. 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 used for model development, researchers must contend with the dual challenges of not overfitting to noise and not underestimating model performance. While existing work mainly highlights the former challenge (*63*), we believe the latter is perhaps more fundamental—and underappreciated.

Noisy training versus noisy test data

Studies in the broader deep learning domain have demonstrated how models trained on noisy but numerate labels can still perform well when evaluated on high-quality test data (*64*–*67*). In sustainable development settings, although noisy training can certainly still degrade model performance when the amount of training data is limited (*68*) or errors are nonrandom, recent studies in agriculture and infrastructure highlight how such noise can be overcome by training on large noisy datasets and/or evaluating on high-quality test data (*53*, *69*, *70*). For example, a satellite-based crop classification model trained on labels derived from millions of imperfectly geolocated smartphone photos in India was able to exceed the performance of benchmark satellite-based classifiers (*69*).

To further explore this ability to overcome noise, we used data from an earlier study of African asset wealth (*28*) to explore the influence on model performance of three types of errors common in publicly available training data: (i) random noise ("jitter") purposely added to village geocoordinates to protect privacy, (ii) sampling variability noise due to small samples, and (iii) noise from household misreporting. We trained models with each type of noise added and evaluated performance on the remaining test data that had either been similarly degraded or unaltered. When trained and evaluated on noisy data, model performance degraded with added noise (Fig. 4, A to C). Yet when evaluated on undegraded test data, model performance remained highly stable, even given large amounts 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 resulting from 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. Training is done on noisier, more numerous data, and testing is done on the sparser high-quality data. A second strategy is to identify an additional variable associated with the outcome of interest, such as weather, in the case of economic output, or fertilizer, in the case of agricultural productivity. The strength of association between this variable and model predictions—as measured, for instance, by correlation—can then be compared with the association between the variable and the (noisy) training data for the model. To illustrate these strategies, Fig. 4, D to F, draws on a recent study of maize yields in Uganda (*53*). Agreement between satellite-based yield estimates and noisy ground data from crop-cuts (i.e., harvests from small, randomly selected portions of a field) has a relatively modest explanatory power (coefficient of determination R^2 = 0.28) (Fig. 4D). Model performance is much better when predictions are compared with the gold-standard measure of full plot harvests, available for a smaller number of randomly selected fields (Fig 4E). Similarly, the correlation between satellite estimates and independent third variables (fertilizer use and soil quality) were the same as the correlation between crop-cut yields and these measures, suggesting that 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 (*18*).

Another example of both strategies is given in (*28*), 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 nonlinear relationships between temperature and wealth, suggesting that the satellite-based wealth measure was roughly as trustworthy as the original ground data.

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 and less devoted to using derived measures for downstream research tasks or policy decisions. 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. We include all published or public preprint studies where comparable test statistics 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.

Smallholder agriculture

Roughly 2.5 billion individuals, and over half of the world's poor, are estimated to live in "smallholder" households that primarily depend on farming small plots of land for their livelihoods (*71*). Although 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 <0.1 ha (or roughly one 30-m Landsat pixel).

Here we assemble data from recent studies attempting to predict yield at the field scale in smallholder environments (table S1), a capability useful for a range of development applications, including the targeting and evaluation of agricultural interventions and the rapid monitoring of rural livelihoods [yield prediction performance at more-aggregate scales is reviewed in (*72*)]. We found 11 studies with comparable field-scale performance metrics, spanning multiple continents and seven crops. All studies used relatively simple models to relate handcrafted 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. Although predictive performance differed widely across and within crops (Fig. 5A), likely owing to the enormous temporal and spatial heterogeneity present in smallholder agriculture, our reanalysis of multiple studies for which replication data were available allowed insight into the determinants of model performance.

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), again highlighting the importance of ground-based measurement error in model evaluation. Also, model performance was much higher on larger fields (Fig. 5C), likely because the same magnitude error can be more consequential for smaller fields; a 10-m georeferencing error is more consequential for a 10-m-wide field than it is for a 100-m-wide field. Finally, additional training samples

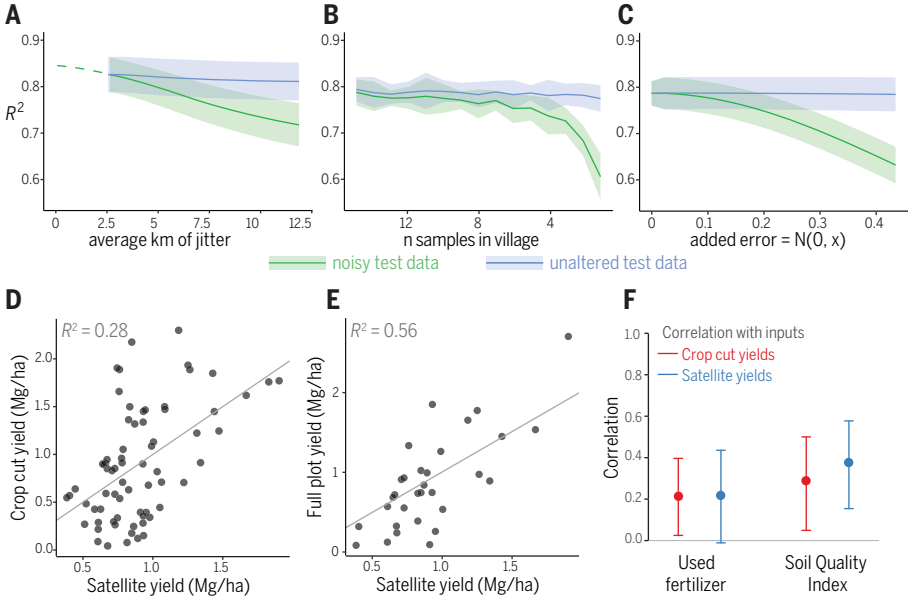


Fig. 4. The role of noise in model performance and evaluation. (**A** to **C**) Performance of wealth prediction model as noise is added to train and test data. Model trained to predict asset wealth from nightlights imagery across 4000 African villages, using the dataset from (*28*). Performance is evaluated as three different types of noise are added to training data: (A) random noise in village geocoordinates (starting from 2.5 km, 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** to **F**) Example from a study of maize yields in Uganda (*53*) 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.

rapidly improved performance on held-out test data (as measured by root mean square error) (Fig. 5D), up to around 30 to 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.

Population

Accurate knowledge of where people are physically located is a critical input into an immense range of research and policy applications. Because population censuses are infrequent in many developing countries and fine-scale data from existing censuses are often not made public, generating fine-scale estimates of population locations has been a research focus for decades.

The traditional top-down approach to local-level population estimation uses satellites and other inputs to redistribute available aggregate census data down to a finer-scale grid (1 km or finer), typically using a model trained on the available coarse-scale data (*73*, *74*). Another approach generates a binary population mask at fine scale using estimates of building or settlement locations derived from imagery and applies this mask to the coarse-scale data (*75*). For either approach, predictions can only be readily evaluated at coarse scale. 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).

For additional quantitative comparison, we studied three commonly used population rasters that used satellite data as at least one input in their production: WorldPop (*74*), Global Human Settlement Layer (GHSL) (*75*), and LandScan (*73*). We harmonized each to a consistent 1-km grid and compared population estimates for grid cells with nonzero estimates across all three rasters. Estimates showed modest agreement (r = 0.62 to 0.78) when comparing across all global pixels (Fig. 6), with

lowest agreement between LandScan and the other rasters and lower overall agreement in Africa (r = 0.45), perhaps owing to limited census data on which to train models. Agreement improves when comparisons are made at increasingly aggregate levels (Fig. 6E), with correlations approaching r = 1.0 when estimates are aggregated to 100-km pixels.

Validation studies in settings where fine-scale population data are available found similar correlations among datasets, e.g., cell-wise correlations between the admin data and GHSL, WorldPop, and LandScan of r = 0.83, 0.82, and 0.7, respectively, in Sweden (*76*). 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*).

Because a standard approach 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, owing to the most recent census having occurred a decade or more prior. An alternative that does not present this problem is to train bottom-up models to directly predict local-level population estimates. These approaches have shown promise in multiple settings (*10*, *24*, *79*), 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 of generating reliable national-scale population estimates (*10*).

Economic livelihoods

Predicting variation in local-level economic outcomes is another active domain, 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 focused on 12 studies that used imagery—either alone or in combination with other data—to predict asset wealth at a local level in the developing world. Asset wealth is a commonly used measure of households' longer-run economic well-being and is consistently measured in a number of georeferenced nationally representative household surveys. Figure 6F shows existing estimates, all of which derive from studies that 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 remained challenging owing to the varied geographic settings (spanning Africa, Asia, and the Caribbean), spatial scales (from village level to district level), and varying inclusion of nonsatellite data, results allowed some generalizations. First, satellite information 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. Again, these estimates likely understate true model performance, as test data almost always derive from public data with known sources of noise. Second, 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 (*27*), Wikipedia (*48*), social media (*49*), or Open Street Map (*51*) can improve predictions.

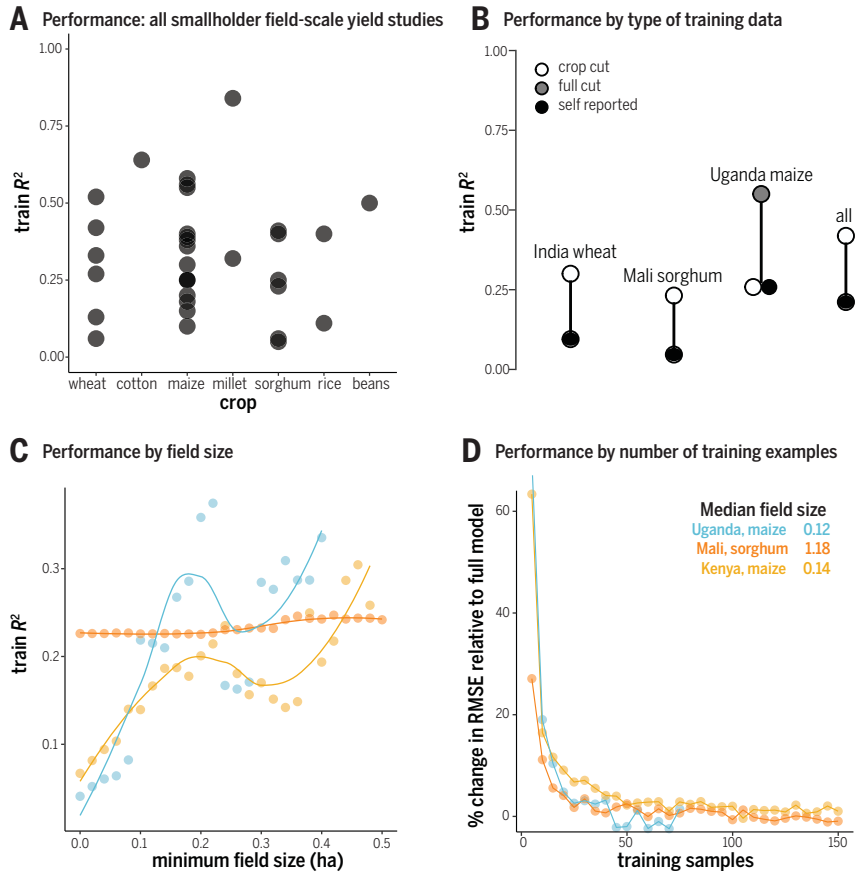


Fig. 5. Performance of satellite-based approaches to measuring smallholder yield at field scale. (**A**) Performance across all known published studies where coefficient of determination (R^2) was reported (32 estimates across 11 studies); R^2 estimates are "in-sample," i.e., for data on which model 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 ~30 data points, and then improves more gradually thereafter. Performance measured as average root mean square 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.

Table S2 describes results from additional studies that looked at other measures of economic livelihoods, including consumption expenditure and multidimensional poverty indices. Prediction performance for consumption expenditure (the measure on which official poverty estimates are based) was typically lower than that for asset wealth, a difference that has been in part attributed to relatively higher noise in the consumption data (*25*, *28*) and the extreme paucity of public georeferenced public consumption 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

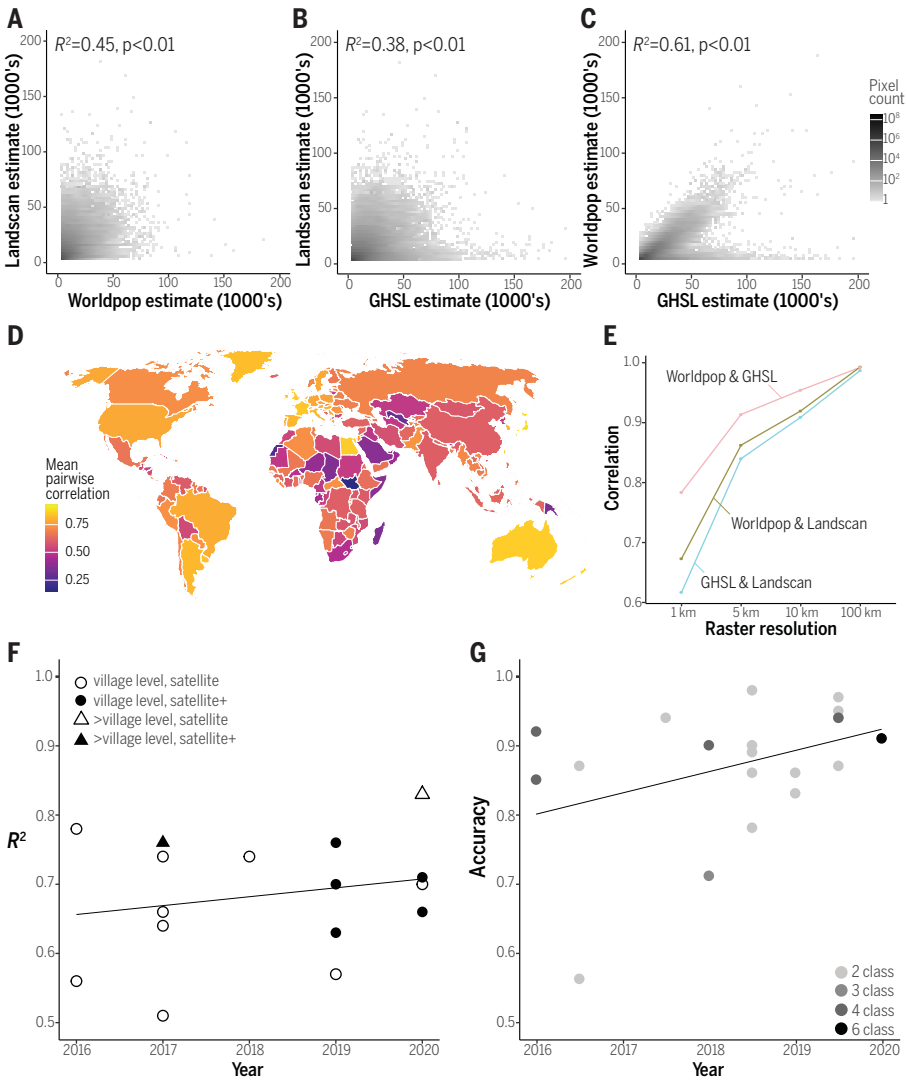


Fig. 6. Performance of satellite-based approaches to measuring population, wealth, and informal settlements. (**A** to **C**) Comparison of three different satellite-informed global population datasets (Landscan, WorldPop, and GHSL population) datasets at 1-km resolution globally (colors correspond to scale at right). (**D**) Average pairwise correlation within each country at 1-km 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 R^2 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 a more-aggregate spatial scale (subdistrict or district). (**G**) Performance in predicting the location of informal settlements from imagery (20 estimates from 17 papers). Colors correspond to the number of categories being predicted.

settlements (sometimes called slums). Urban populations are growing rapidly throughout much of the developing world, and ~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 (*8*).

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 (*8*) 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.

As with the other domains discussed, studies span diverse geographies where 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.

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 public health, disaster response, economic development, and climate change research (*10*, *80*, *82*). Satellite imagery has also been widely used to better understand agricultural productivity, including why some fields or some regions are more productive than others (*6*), whether particular management practices have been adopted (*83*), and the impact of infrastructure investments on productivity (*84*, *85*). Satellite estimates are also increasingly being used to identify fields most likely to respond to a particular input (*18*, *53*) or new management practice (*86*).

Fisheries and animal production are additional food-related domains where satellite imagery is 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 (*35*) as well as on specific activities, such as illegal fishing (*36*, *87*).

Researchers in economics also increasingly use satellite imagery, and particularly nightlights imagery, for a variety of applications (*7*). Nightlights have been used to assess the validity of official government statistics (*19*, *88*); to understand the growth and activity of urban versus rural areas (*89*, *90*); and to assess the role of local and federal institutions, transport costs, and other factors on economic development (*91*–*94*). 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 (*95*) and to understand how ethnic favoritism shapes economic investment (*96*).

Recent work also shows how satellites can be useful in the experimental evaluation of interventions, including measuring the heterogeneous impact of new agricultural technologies on productivity (*86*), measuring the impacts of cash transfers on household livelihoods (*97*), and measuring compliance in a payment-for-forest-protection program (*95*). While each 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.

Use in decision-making

Although 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. Thus, imagery has so far done more to help understand sustainable development, and less to promote it. Systematic information on operational use in the private or military sector is even more sparse, although use is likely widespread and growing. Here we only consider public-sector nonmilitary use.

As in research, the widest application of satellite-based measures in public-sector decision-making is in the population domain. For instance, the United Nations World Food Programme and the US government both use gridded population estimates to inform needs assessments and target humanitarian response after 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 (*98*). Numerous systems that track agricultural conditions 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 trade policy (*99*). Data from remote detection of fishing activity is also being used by numerous governments and other organizations to manage fisheries and design protected areas (*100*).

Documented use in other livelihoods measurement appears limited, although anecdotally there is rapidly growing interest in the policy community in exploring these measures (*101*). The simplest explanation for limited adoption is that the combination of satellite information and machine learning is still new and decision-makers are unfamiliar with these approaches or are not convinced that 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, 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 (*102*)] and concerns by civil rights groups that further deployment of algorithmic decision-making might worsen racial and socioeconomic inequalities (*103*, *104*) 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) (*105*), if followed, could help navigate these issues. These guidelines aim to ensure that researchers are aware of potential discriminatory effects of their algorithms and are able to investigate and provide redress should issues arise. While implementation of the guidelines certainly has its own challenges (*106*) (e.g., defining "fairness"), we are not aware of any of the papers we review above—our own included—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 (Fig. 1), and certain countries have passed laws (since reversed) that make it a crime to publish independent estimates of key economic outcomes (*107*).

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 with 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 will 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 constitutes a key task for researchers and policymakers going forward. We suggest nine specific

areas where we believe future work would be particularly useful:

1) More accurate and more numerous training data: Many applications of deep learning outside sustainable development have been advanced by the curation of public reference datasets, which lower barriers to entry and enable comparison of different approaches. Such datasets are a major public good but are rare in sustainable development. Particularly needed are datasets that track outcomes over time so that models can be optimized to detect changes. Collecting and publishing location data from existing or new ground surveys (using appropriate privacy safeguards already widely in use) could be mandated by survey funders.

2) More evaluation in the context of specific use cases: Most evaluations 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 subpopulation. These downstream tasks often provide a more tangible example of the utility to potential users and can help avoid the pitfalls of direct comparisons to noisy ground measures.

3) Improved model interpretability and transparency: Interpretable predictions and transparent decisions based on these predictions are important in settings where people could be affected. Applying FAT ML or similar guidelines to research output will also be important as research is operationalized.

4) Creative data fusion: Combining information from optical sensors of different temporal and spatial resolutions, different types of imagery (e.g., optical and radar), and/or alternate data streams (e.g., from cell phones) appears to be a particularly promising approach 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 (*108*).

5) Scaling estimates: Researchers typically have more incentive to innovate on methods than to apply validated methods across large geographies or to update estimates as new data come in, limiting the utility of methodological advances to downstream use. Partnerships between academic researchers and public- or private-sector organizations that have the skills and resources to do this scaling will be key to operationalizing promising research advances.

6) Measuring changes over time: Much of the literature reviewed above makes predictions at a given point in time, but many applications require measuring changes over time. As few ground datasets 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: Improved satellite predictions 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 whether such sampling strategies could improve outcome measurement as compared with 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 are 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 AND NOTES

1. G. K. Moore, What is a picture worth? A history of remote sensing. *Hydrol. Sci. Bull.* **24**, 477–485 (1979). doi: 10.1080/02626667909491887
2. O. B. Waxman, "Aerial photography's surprising role in history," *Time Magazine,* 31 May 2018; https://time.com/longform/aerial-photography-drones-history/.
3. Union of Concerned Scientists, UCS Satellite Database (2020); www.ucsusa.org/resources/satellite-database.
4. United Nations General Assembly, "Transforming our world: The 2030 Agenda for Sustainable Development" (Division for Sustainable Development Goals, 2015); https://sdgs.un.org/2030agenda.
5. D. J. Mulla, Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosyst. Eng.* **114**, 358–371 (2013). doi: 10.1016/j.biosystemseng.2012.08.009
6. D. B. Lobell, The use of satellite data for crop yield gap analysis. *Field Crops Res.* **143**, 56–64 (2013). doi: 10.1016/j.fcr.2012.08.008
7. D. Donaldson, A. Storeygard, The view from above: Applications of satellite data in economics. *J. Econ. Perspect.* **30**, 171–198 (2016). doi: 10.1257/jep.30.4.171
8. M. Kuffer, K. Pfeffer, R. Sliuzas, Slums from space—15 years of slum mapping using remote sensing. *Remote Sens.* **8**, 455–484 (2016). doi: 10.3390/rs8060455
9. Sustainable Development Solutions Network, "Data for development: A needs assessment for SDG monitoring and statistical capacity development" (UN Sustainable Development Solutions Network, 2015).
10. N. A. Wardrop *et al*., Spatially disaggregated population estimates in the absence of national population and housing census data. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3529–3537 (2018). doi: 10.1073/pnas.1715305115; pmid: 29555739
11. S. Devarajan, Africa's statistical tragedy. *Rev. Income Wealth* **59**, S9–S15 (2013). doi: 10.1111/roiw.12013
12. K. Beegle, J. De Weerdt, J. Friedman, J. Gibson, Methods of household consumption measurement through surveys: Experimental results from Tanzania. *J. Dev. Econ.* **98**, 3–18 (2012). doi: 10.1016/j.jdeveco.2011.11.001
13. C. Carletto, D. Jolliffe, R. Banerjee, From tragedy to renaissance: Improving agricultural data for better policies. *J. Dev. Stud.* **51**, 133–148 (2015). doi: 10.1080/00220388.2014.968140
14. "Capacity needs assessment for improving agricultural statistics in Kenya," (Tech. rep., The World Bank, 2018).
15. R. B. Macdonald, F. G. Hall, Global crop forecasting. *Science* **208**, 670–679 (1980). doi: 10.1126/science.208.4445.670; pmid: 17771086
16. C. D. Elvidge *et al*., Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *Int. J. Remote Sens.* **18**, 1373–1379 (1997). doi: 10.1080/014311697218485
17. See supplementary materials.
18. M. Burke, D. B. Lobell, Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 2189–2194 (2017). doi: 10.1073/pnas.1616919114; pmid: 28202728
19. J. V. Henderson, A. Storeygard, D. N. Weil, Measuring economic growth from outer space. *Am. Econ. Rev.* **102**, 994–1028 (2012). doi: 10.1257/aer.102.2.994; pmid: 25067841
20. M. Weiss, F. Jacob, G. Duveiller, Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* **236**, 111402–111421 (2020). doi: 10.1016/j.rse.2019.111402
21. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2016**, 770–778 (2016).
22. T. G. Tiecke *et al*., Mapping the world population one building at a time. arXiv:1712.05839 [cs.CV] (15 December 2017).
23. Z. Zong, J. Feng, K. Liu, H. Shi, Y. Li, DeepDPM: Dynamic Population Mapping via Deep Neural Network. *Proc. Conf. AAAI Artif. Intell.* **33**, 1294–1301 (2019). doi: 10.1609/aaai.v33i01.33011294
24. W. Hu *et al*., "Mapping missing population in rural India: A deep learning approach with satellite imagery," *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Honolulu, Hawaii, 27 to 28 January 2019, pp. 353–359.
25. N. Jean *et al*., Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016). doi: 10.1126/science.aaf7894; pmid: 27540167
26. A. Head, M. Manguin, N. Tran, J. E. Blumenstock, "Can human development be measured with satellite imagery?" *ICTD 2017: Ninth International Conference on Information and Communication Technologies and Development*, Lahore, Pakistan, 16 to 19 November 2017.
27. J. E. Steele *et al*., Mapping poverty using mobile phone and satellite data. *J. R. Soc. Interface* **14**, 20160690–20160700 (2017). doi: 10.1098/rsif.2016.0690; pmid: 28148765
28. C. Yeh *et al*., Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nat. Commun.* **11**, 2583 (2020). doi: 10.1038/s41467-020-16185-w; pmid: 32444658
29. G. Cadamuro, A. Muhebwa, J. Taneja, Assigning a grade: Accurate measurement of road quality using satellite imagery. arXiv:1812.01699 [cs.CV] (6 December 2018).
30. B. Oshri *et al*., "Infrastructure Quality Assessment in Africa using Satellite Imagery and Deep Learning," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, London, UK, 19 to 23 August 2018.
31. A. Albert, J. Kaur, M. C. Gonzalez, "Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Nova Scotia, Canada, 13 to 17 August 2017, pp. 1357–1366.
32. P. Helber, B. Bischke, A. Dengel, D. Borth, Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **12**, 2217–2226 (2019). doi: 10.1109/JSTARS.2019.2918242
33. N. Mboga, C. Persello, J. R. Bergado, A. Stein, Detection of informal settlements from VHR images using convolutional neural networks. *Remote Sens.* **9**, 1106–1124 (2017). doi: 10.3390/rs9111106
34. C. Persello, A. Stein, Deep fully convolutional networks for the detection of informal settlements in VHR images. *IEEE Geosci. Remote Sens. Lett.* **14**, 2325–2329 (2017). doi: 10.1109/LGRS.2017.2763738

35. D. A. Kroodsma *et al*., Tracking the global footprint of fisheries. *Science* **359**, 904–908 (2018). doi: 10.1126/science.aao5646; pmid: 29472481
36. J. Park *et al*., Illuminating dark fishing fleets in North Korea. *Sci. Adv.* **6**, eabb1197 (2020). doi: 10.1126/sciadv.abb1197; pmid: 32923605
37. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997). doi: 10.1162/neco.1997.9.8.1735; pmid: 9377276
38. X. Shi *et al*., Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **28**, 802–810 (2015).
39. S. Ji, C. Zhang, A. Xu, Y. Shi, Y. Duan, 3D convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sens.* **10**, 75–92 (2018). doi: 10.3390/rs10010075
40. M. Rußwurm, M. Körner, Multi-temporal land cover classification with long short-term memory neural networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci* **42**, 551–558 (2017). doi: 10.5194/isprs-archives-XLII-1-W1-551-2017
41. R. M. Rustowicz *et al*., Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2019**, 75–82 (2019).
42. J. You, X. Li, M. Low, D. Lobell, S. Ermon, Deep gaussian process for crop yield prediction based on remote sensing data. *Proc. Conf. AAAI Artif. Intell* **31**, 4559–4566 (2017).
43. J. Sun, L. Di, Z. Sun, Y. Shen, Z. Lai, County-level soybean yield prediction using deep CNN-LSTM model. *Sensors* **19**, 4363 (2019). doi: 10.3390/s19204363; pmid: 31600963
44. L. Xiao, Y. Zhang, G. Peng, Landslide susceptibility assessment using integrated deep learning algorithm along the China-Nepal highway. *Sensors* **18**, 4436 (2018). doi: 10.3390/s18124436; pmid: 30558225
45. J. Z. Xu, W. Lu, Z. Li, P. Khaitan, V. Zaytseva, Building damage detection in satellite imagery using convolutional neural networks. arXiv:1910.06444 [cs.CV] (14 October 2019).
46. T. Ci, Z. Liu, Y. Wang, Assessment of the degree of building damage caused by disaster using convolutional neural networks in combination with ordinal regression. *Remote Sens.* **11**, 2858–2877 (2019). doi: 10.3390/rs11232858
47. F. M. Davenport *et al*., Using out-of-sample yield forecast experiments to evaluate which earth observation products best indicate end of season maize yields. *Environ. Res. Lett.* **14**, 124095–124109 (2019). doi: 10.1088/1748-9326/ab5ccd
48. E. Sheehan *et al*., "Predicting economic development using geolocated Wikipedia articles," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, Alaska, 4 to 8 August 2019, pp. 2698–2706.
49. M. Fatehkia, B. Coles, F. Ofli, I. Weber, "The Relative Value of Facebook Advertising Data for Poverty Mapping, *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, Atlanta, Georgia, 8 to 11 June 2019, pp. 934–938.
50. R. Cao *et al*., Integrating aerial and street view images for urban land use classification. *Remote Sens.* **10**, 1553–1576 (2018). doi: 10.3390/rs10101553
51. I. Tingzon *et al*., "Mapping Poverty in the Philippines Using Machine Learning, Satellite Imagery, and Crowd-sourced Geospatial Information," *International Conference on Machine Learning AI for Social Good Workshop*, Long Beach, California, 15 June 2019.
52. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, 21 to 26 July 2017, pp. 2261–2269.
53. D. B. Lobell *et al*., Eyes in the sky, boots on the ground: Assessing satellite-and ground-based approaches to crop yield measurement and analysis. *Am. J. Agric. Econ.* **102**, 202–219 (2020). doi: 10.1093/ajae/aaz051
54. G. Christie, N. Fendley, J. Wilson, R. Mukherjee, Functional Map of the World. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2018**, 6172–6180 (2018).
55. B. Uzkent *et al*., "Learning to interpret satellite images using Wikipedia," *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, 10 to 16 August 2019, pp. 3620–3626.
56. K. Ayush, B. Uzkent, M. Burke, D. Lobell, S. Ermon, "Generating interpretable poverty maps using object detection in satellite images," *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, January 2021, pp. 4410–4416.
57. K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2020**, 9729–9738 (2020).
58. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations. arXiv:2002.05709 [cs.LG] (1 July 2020).
59. S. Basu *et al*., "Deepsat: a learning framework for satellite imagery," *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Seattle, Washington, 3 to 6 November 2015.
60. N. Jean, S. Wang, G. Azzari, D. Lobell, S. Ermon, Tile2Vec: Unsupervised representation learning for spatially distributed data. *Proc. Conf. AAAI Artif. Intell.* **33**, 3967–3974 (2019). doi: 10.1609/aaai.v33i01.33013967
61. E. Rolf *et al*., "A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery" (NBER Working Paper 28045, National Bureau of Economic Research, 2020).
62. N. Jean, S. M. Xie, S. Ermon, Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. *Adv. Neural Inf. Process. Syst.* **31**, 5322–5333 (2018).
63. A. Elmes *et al*., Accounting for training data error in machine learning applied to Earth observations. *Remote Sens.* **12**, 1034–1073 (2020). doi: 10.3390/rs12061034
64. J. Krause *et al*., "The unreasonable effectiveness of noisy data for fine-grained recognition," *14th European Conference on Computer Vision*, Amsterdam, Netherlands, 8 to 16 October 2016.
65. D. Rolnick, A. Veit, S. Belongie, N. Shavit, Deep learning is robust to massive label noise. arXiv:1705.10694 [cs.LG] (26 February 2018).
66. N. Natarajan, I. S. Dhillon, P. K. Ravikumar, A. Tewari, Learning with noisy labels. *Adv. Neural Inf. Process. Syst.* **26**, 1196–1204 (2013).
67. M. Charikar, J. Steinhardt, G. Valiant, "Learning from untrusted data," *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, Montreal, Quebec, Canada, 19 to 23 June 2017, pp. 47–60.
68. A. Paliwal, M. Jain, The accuracy of self-reported crop yield estimates and their ability to train remote sensing algorithms. *Front. Sustain. Food Syst.* **4**, 25–35 (2020). doi: 10.3389/fsufs.2020.00025
69. S. Wang *et al*., Mapping crop types in southeast India with smartphone crowdsourcing and deep learning. *Remote Sens.* **12**, 2957–2999 (2020). doi: 10.3390/rs12182957
70. P. Kaiser *et al*., Learning aerial image segmentation from online maps. *IEEE Trans. Geosci. Remote Sens.* **55**, 6054–6068 (2017). doi: 10.1109/TGRS.2017.2719738
71. R. P. Christen, J. Anderson, "Segmentation of smallholder households: Meeting the range of financial needs in agricultural families," Focus Note 85, CGAP, Washington, DC, April 2013.
72. W. Chivasa, O. Mutanga, C. Biradar, Application of remote sensing in estimating maize grain yield in heterogeneous African agricultural landscapes: A review. *Int. J. Remote Sens.* **38**, 6816–6845 (2017). doi: 10.1080/01431161.2017.1365390
73. J. E. Dobson, E. A. Bright, P. R. Coleman, R. C. Durfee, B. A. Worley, LandScan: A global population database for estimating populations at risk. *Photogramm. Eng. Remote Sensing* **66**, 849–857 (2000).
74. F. R. Stevens, A. E. Gaughan, C. Linard, A. J. Tatem, Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLOS ONE* **10**, e0107042 (2015). doi: 10.1371/journal.pone.0107042; pmid: 25689585
75. M. Schiavina, S. Freire, K. MacManus, GHS-POP R2019A - GHS population grid multitemporal (1975-1990-2000-2015), Dataset, European Commission, Joint Research Centre (JRC) (2019); http://data.europa.eu/89h/0c6b9751-a71f-4062-830b-43c9f432370f.
76. M. F. A. Bustos, O. Hall, T. Niedomysl, U. Ernstson, A pixel level evaluation of five multitemporal global gridded population datasets: A case study in Sweden, 1990–2015. *Popul. Environ.* **42**, 255–277 (2020). doi: 10.1007/s11111-020-00360-8
77. B. Calka, E. Bielecka, GHS-POP accuracy assessment: Poland and Portugal case study. *Remote Sens.* **12**, 1105–1128 (2020). doi: 10.3390/rs12071105
78. Z. Bai, J. Wang, M. Wang, M. Gao, J. Sun, Accuracy assessment of multi-source gridded population distribution datasets in China. *Sustainability* **10**, 1363 (2018). doi: 10.3390/su10051363
79. R. Engstrom, D. Newhouse, V. Soundararajan, Estimating small-area population density in Sri Lanka using surveys and Geo-spatial data. *PLOS ONE* **15**, e0237063 (2020). doi: 10.1371/journal.pone.0237063; pmid: 32756580
80. Thematic Research Network on Data and Statistics, "Leaving no one off the map: A guide for gridded population data for sustainable development," (UN Sustainable Development Solutions Network, 2020).
81. "Habitat III Issue Paper 22–Informal Settlements," United Nations Conference on Housing and Sustainable Urban Development, Quito, Ecuador, 17 to 20 October 2016.
82. S. Leyk *et al*., The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth Syst. Sci. Data* **11**, 1385–1409 (2019). doi: 10.5194/essd-11-1385-2019
83. C. Kubitza, V. V. Krishna, U. Schulthess, M. Jain, Estimating adoption and impacts of agricultural management practices in developing countries using satellite data. A scoping review. *Agron. Sustain. Dev.* **40**, 16 (2020). doi: 10.1007/s13593-020-0610-2
84. E. Strobl, R. O. Strobl, The distributional impact of large dams: Evidence from cropland productivity in Africa. *J. Dev. Econ.* **96**, 432–450 (2011). doi: 10.1016/j.jdeveco.2010.08.005
85. E. Blanc, E. Strobl, Is small better? A comparison of the effect of large and small dams on cropland productivity in South Africa. *World Bank Econ. Rev.* **28**, 545–576 (2014). doi: 10.1093/wber/lht026
86. M. Jain *et al*., The impact of agricultural interventions can be doubled by using satellite data. *Nat. Sustain.* **2**, 931–934 (2019). doi: 10.1038/s41893-019-0396-x
87. D. Belhabib *et al*., Catching industrial fishing incursions into inshore waters of Africa from space. *Fish Fish.* **21**, 379–392 (2020). doi: 10.1111/faf.12436
88. M. Pinkovskiy, X. Sala-i-Martin, Lights, camera… income! Illuminating the national accounts-household surveys debate. *Q. J. Econ.* **131**, 579–631 (2016). doi: 10.1093/qje/qjw003
89. V. Henderson, T. Squires, A. Storeygard, D. Weil, The global distribution of economic activity: Nature, history, and the role of trade. *Q. J. Econ.* **133**, 357–406 (2018). doi: 10.1093/qje/qjx030; pmid: 31798191
90. M. Harari, Cities in bad shape: Urban geometry in India. *Am. Econ. Rev.* **110**, 2377–2421 (2020). doi: 10.1257/aer.20171673
91. S. Michalopoulos, E. Papaioannou, Pre-colonial ethnic institutions and contemporary African development. *Econometrica* **81**, 113–152 (2013). doi: 10.3982/ECTA9613; pmid: 25089052
92. S. Michalopoulos, E. Papaioannou, National institutions and subnational development in Africa. *Q. J. Econ.* **129**, 151–213 (2013). doi: 10.1093/qje/qjt029; pmid: 25802926
93. A. Storeygard, Farther on down the road: Transport costs, trade and urban growth in sub-Saharan Africa. *Rev. Econ. Stud.* **83**, 1263–1295 (2016). doi: 10.1093/restud/rdw020; pmid: 29743731
94. M. L. Pinkovskiy, Growth discontinuities at borders. *J. Econ. Growth* **22**, 145–192 (2017). doi: 10.1007/s10887-016-9139-2
95. S. Jayachandran *et al*., Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science* **357**, 267–273 (2017). doi: 10.1126/science.aan0568; pmid: 28729505
96. B. Marx, T. M. Stoker, T. Suri, There is no free house: Ethnic patronage in a Kenyan slum. *Am. Econ. J. Appl. Econ.* **11**, 36–70 (2019). doi: 10.1257/app.20160484
97. L. Y. Huang, "Measuring the impacts of poverty alleviation programs with satellite imagery and deep learning" (2020); http://luna-yue-huang.com/assets/pdf/jmp.pdf.
98. M. E. Brown, *Famine Early Warning Systems and Remote Sensing Data* (Springer Science & Business Media, 2008).
99. S. Fritz *et al*., A comparison of global agricultural monitoring systems and current gaps. *Agric. Syst.* **168**, 258–272 (2019). doi: 10.1016/j.agsy.2018.05.010
100. Global Fishing Watch, Ocean sustainability through transparency, data-sharing and collaboration (2020); https://globalfishingwatch.org/wp-content/uploads/GFW-program-2020.pdf.
101. J. Blumenstock, Machine learning can help get COVID-19 aid to those who need it most. *Nature* 10.1038/d41586-020-01393-7 (2020). doi: 10.1038/d41586-020-01393-7; pmid: 32409767

102. D. Cossins, "Discriminating algorithms: 5 times AI showed prejudice," *New Scientist*, 12 April 2018; www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/.
103. Data for Black Lives, https://d4bl.org.
104. The Leadership Conference on Civil and Human Rights, Civil rights principles for the era of big data; https://civilrights.org/2014/02/27/civil-rights-principles-era-big-data/.
105. N. Diakopoulos *et al*., Principles for Accountable Algorithms and a Social Impact Statement for Algorithms; www.fatml.org/resources/principles-for-accountable-algorithms.
106. P. Gajane, M. Pechenizkiy, On formalizing fairness in prediction with machine learning. arXiv:1710.03184 [cs.LG] (28 May 2018).
107. O. Nyeko, "Tanzania drops threat of prison over publishing independent statistics: Amendment to statistics act a step in right direction for free expression," Human Rights Watch, 3 July 2019; www.hrw.org/news/2019/07/03/tanzania-drops-threat-prison-over-publishing-independent-statistics.
108. D. M. J. Lazer *et al*., Computational social science: Obstacles and opportunities. *Science* **369**, 1060–1062 (2020). doi: 10.1126/science.aaz8170; pmid: 32855329
109. F. Solt, The Standardized World Income Inequality Database, Version 8, Harvard Dataverse (2019).
110. The World Bank, World Development Indicators (2020); http://wdi.worldbank.org.
111. M. G. Marshall, T. R. Gurr, POLITY5: Political Regime Characteristics and Transitions, 1800-2018 (Center for Systemic Peace, 2020); www.systemicpeace.org/inscr/p5manualv2018.pdf.
112. A. Driscoll, burke-lab/satellite-review-public: Initial release of data and code, Version 1, Zenodo (2021); http://doi.org/10.5281/zenodo.4417632.

ACKNOWLEDGMENTS

We thank J. Xue, B. Lin, and Z. Tang for excellent research assistance. **Funding:** We thank USAID Bureau for Food Security, the Global Innovation Fund, Darpa World Modelers program, NSF grants 1651565 and 1522054, and the Stanford King Center on Global Development for funding. **Competing interests:** M.B., D.B.L., and S.E. are co-founders of AtlasAI, a company that uses machine learning to measure economic outcomes in the developing world. **Data and materials availability:** Data and code for replication of all results are available at https://github.com/burke-lab/satellite-review and (*112*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6535/eabe8628/suppl/DC1
Materials and Methods
Tables S1 to S3
References (*113*–*155*)

10.1126/science.abe8628