
SUSTAINBENCH: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning

Christopher Yeh* **Chenlin Meng*** **Sherrie Wang*** **Anne Driscoll†** **Erik Rozi†**
Caltech Stanford UC Berkeley Stanford Stanford

Patrick Liu† **Jihyeon Lee†** **Marshall Burke** **David Lobell** **Stefano Ermon**
Stanford Stanford Stanford Stanford Stanford

Abstract

Progress toward the United Nations Sustainable Development Goals (SDGs) has been hindered by a lack of data on key environmental and socioeconomic indicators, which historically have come from ground surveys with sparse temporal and spatial coverage. Recent advances in machine learning have made it possible to utilize abundant, frequently-updated, and globally available data, such as from satellites or social media, to provide insights into progress toward SDGs. Despite promising early results, approaches to using such data for SDG measurement thus far have largely evaluated on different datasets or used inconsistent evaluation metrics, making it hard to understand whether performance is improving and where additional research would be most fruitful. Furthermore, processing satellite and ground survey data requires domain knowledge that many in the machine learning community lack. In this paper, we introduce SUSTAINBENCH, a collection of 15 benchmark tasks across 7 SDGs, including tasks related to economic development, agriculture, health, education, water and sanitation, climate action, and life on land. Datasets for 11 of the 15 tasks are released publicly for the first time. Our goals for SUSTAINBENCH are to (1) lower the barriers to entry for the machine learning community to contribute to measuring and achieving the SDGs; (2) provide standard benchmarks for evaluating machine learning models on tasks across a variety of SDGs; and (3) encourage the development of novel machine learning methods where improved model performance facilitates progress towards the SDGs.

1 Introduction

In 2015, the United Nations (UN) proposed 17 Sustainable Development Goals (SDGs) to be achieved by 2030, for promoting prosperity while protecting the planet [2]. The SDGs span social, economic, and environmental spheres, ranging from ending poverty to achieving gender equality to combating climate change (see Table A1). Progress toward SDGs is traditionally monitored through statistics collected by civil registrations, population-based surveys and censuses. However, such data collection is expensive and requires adequate statistical capacity, and many countries go decades between making ground measurements on key SDG indicators [20]. Only roughly half of SDG indicators have regular data from more than half of the world’s countries [94]. These data gaps severely limit the ability of the international community to track progress toward the SDGs.

*Joint first authors.

†Joint second authors.

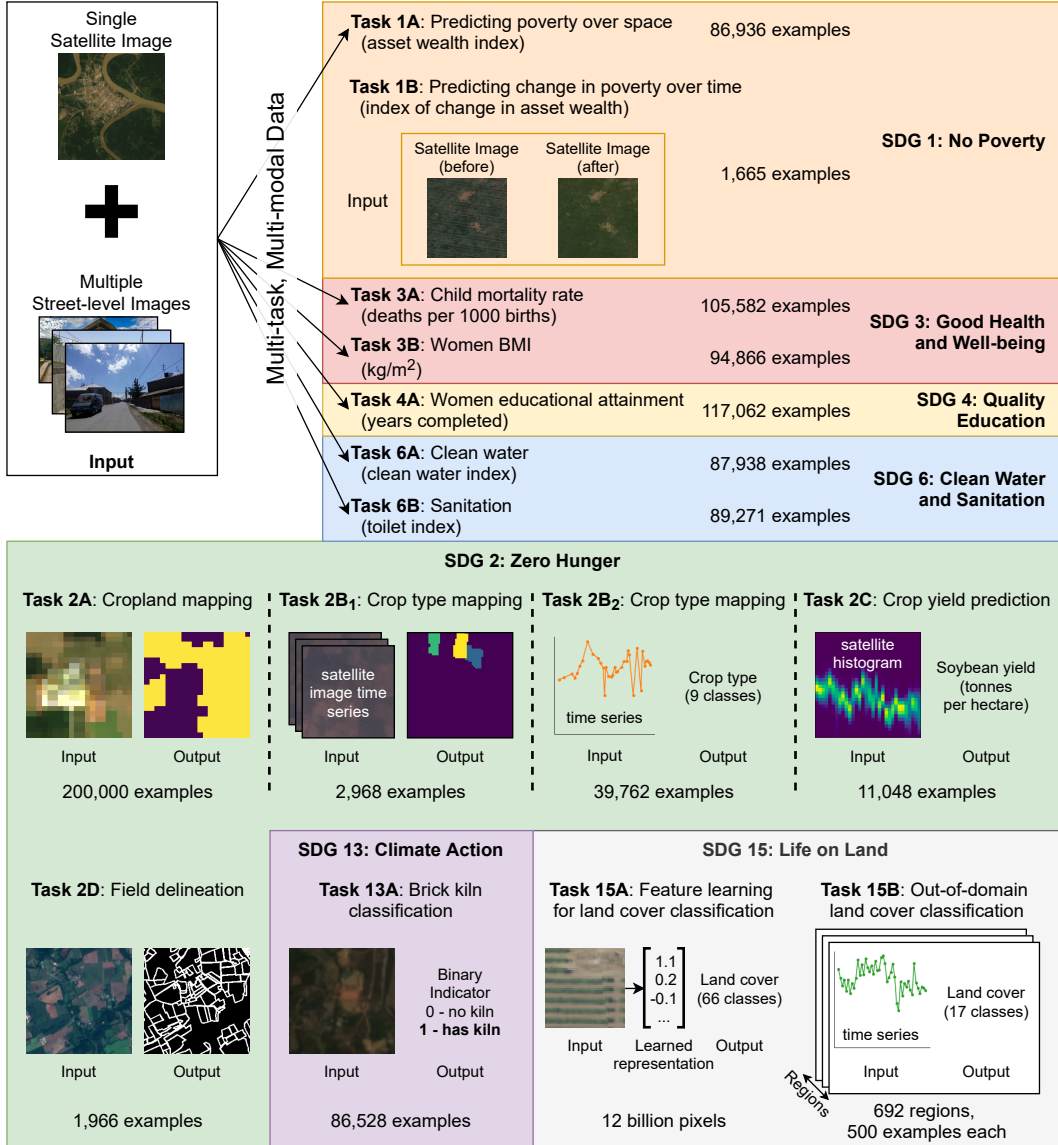


Figure 1: Datasets and tasks included in SUSTAINBENCH ranging from poverty prediction to land cover classification (described in Section 3 with additional details in Appendix D). Data for 11 out of 15 tasks are publicly released for the first time.

Advances in machine learning (ML) have shown promise in helping plug these data gaps, demonstrating how sparse ground data can be combined with abundant, cheap and frequently updated sources of novel sensor data to measure a range of SDG-related outcomes [70, 20]. For instance, data from satellite imagery, social media posts, and/or mobile phone activity can predict poverty [15, 52, 109], annual land cover [35, 18], deforestation [42, 50], agricultural cropping patterns [69, 103], crop yields [11, 110], and the location and impact of natural disasters [25, 92]. As a timely example of real-world impact, the governments of Bangladesh, Mozambique, Nigeria, Togo, and Uganda used ML-based poverty and cropland maps generated from satellite imagery or phone records to target economic aid to their most vulnerable populations during the COVID-19 pandemic [14, 38, 56, 66]. Other recent work demonstrates using ML-based poverty maps to measure the effectiveness of large-scale infrastructure investments [78].

But further methodological progress on the “big data approach” to monitoring SDGs is hindered by a number of key challenges. First, downloading and working with both novel input data (e.g., from satellites) and ground-based household surveys requires domain knowledge that many in the ML community lack. Second, existing approaches have been evaluated on different datasets, data splits,

or evaluation metrics, making it hard to understand whether performance is improving and where additional research would be most fruitful [20]. This is in stark contrast to canonical ML datasets like MNIST, CIFAR-10 [60], and ImageNet [81] that have standardized inputs, outputs, and evaluation criteria and have therefore facilitated remarkable algorithmic advances [43, 28, 57, 44, 47]. Third, methods used so far are often adapted from methods originally designed for canonical deep learning datasets (*e.g.*, ImageNet). However, the datasets and tasks relevant to SDGs are unique enough to merit their own methodology. For example, gaps in monitoring SDGs are widest in low-income countries, where only sparse ground labels are available to train or validate predictive models.

To facilitate methodological progress, this paper presents SUSTAINBENCH, a compilation of datasets and benchmarks for monitoring the SDGs with machine learning. Our goals are to

1. lower the barriers to entry by supplying high-quality domain-specific datasets in development economics and environmental science,
2. provide benchmarks to standardize evaluation on tasks related to SDG monitoring, and
3. encourage the ML community to evaluate and develop novel methods on problems of global significance where improved model performance facilitates progress towards SDGs.

In SUSTAINBENCH, we curate a suite of 15 benchmark tasks across 7 SDGs where we have relatively high-quality ground truth labels: No Poverty (SDG 1), Zero Hunger (SDG 2), Good Health and Well-being (SDG 3), Quality Education (SDG 4), Clean Water and Sanitation (SDG 6), Climate Action (SDG 13), and Life on Land (SDG 15). Figure 1 summarizes the datasets in SUSTAINBENCH. Although results for some tasks have been published previously, *data for 11 of the 15 tasks are being made public for the first time*. We provide baseline models for each task and a public leaderboard³.

To our knowledge, this is the first set of large-scale cross-domain datasets targeted at SDG monitoring compiled with standardized data splits to enable benchmarking. SUSTAINBENCH is not only valuable to improving sustainability measurements but also offers tasks for ML challenges, allowing for the development of self-supervised learning (Section 3.7), meta-learning (Section 3.7), and multi-modal/multi-task learning methods (Sections 3.1 and 3.3 to 3.5) on real-world datasets.

In the remainder of this paper, Section 2 surveys related datasets; Section 3 introduces the SDGs and datasets covered by SUSTAINBENCH; Section 4 summarizes state-of-the-art models on each dataset and where methodological advances are needed; and Section 5 highlights the impact, limitations, and future directions of this work. The Appendix includes detailed information about the inputs, labels, and tasks for each dataset.

2 Related Work

Our work builds on a growing body of research that seeks to measure SDG-relevant indicators, including those cited above. These individual studies typically focus on only one SDG-related task, but even within a specific SDG domain (*e.g.*, poverty prediction), most tasks lack standardized datasets with clear replicate-able benchmarks [20]. In comparison, SUSTAINBENCH is a compilation of datasets that covers 7 SDGs and provides 15 standardized, replicate-able tasks with established benchmarks. Table 1 compares SUSTAINBENCH against existing datasets that pertain to SDGs, are publicly available, provide ML-friendly inputs/outputs, and specify standardized evaluation metrics.

Perhaps the most closely-related benchmark dataset is WILDS [59], which provides a comprehensive benchmark for distribution shifts in real-world applications. However, WILDS is not focused on SDGs, and although it includes a poverty mapping task, our poverty dataset covers $5\times$ more countries.

There also exist a number of datasets for performing satellite or aerial imagery tasks related to the SDGs [23, 86, 89, 108, 96, 62, 41, 4, 26, 96] which share similarities with the inputs of SUSTAINBENCH on certain benchmarks. For example, [86] compiled imagery from the Sentinel-1/2 satellites, which we also use for SDG monitoring tasks, and the Radiant Earth Foundation has compiled datasets for crop type mapping [77], a task we also include. However, SUSTAINBENCH’s goal is to provide a broader view of what ML can do for SDG monitoring; it is differentiated in its focus on multiple SDGs, multiple inputs, and on low-income regions in particular. For tasks where existing datasets are abundant (*e.g.*, cropland and land cover classification), SUSTAINBENCH has tasks that address

³<https://sustainlab-group.github.io/sustainbench/leaderboard>

Table 1: A comparison of SUSTAINBENCH with related datasets and benchmarks. A dataset is only included if it is relevant for an SDG, is publicly available, provides both inputs and outputs in ML-friendly formats, defines train/test sets, and standardizes evaluation metrics.

Name	Purpose	Geography	Time	Inputs	Relevant for SDGs									
					1	2	3	4	6	11	13	14	15	
SUSTAINBENCH	SDG monitoring	1-105 countries/task (119 total)	1-24 years/task in 1996-2019	Sat. images, street-level images, and/or time series	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Yeh <i>et al.</i> / WILDS [109, 59]	Poverty mapping	23 countries	2009-16	Sat. images	✓									
Radiant MLHub [77]	Crop type mapping	8 countries	1-3 years/task in 2015-21	Sat. time series or drone images		✓								
SpaceNet [96]	Building & road detection	10+ cities	Unknown	Sat. images & time series						✓				
DeepGlobe [26]	Building & road detection, land cover classification	3 countries, 4 cities	Unknown	Sat. images						✓			✓	
fMoW / WILDS [23, 59]	Object detection	207 countries	2002-17	Sat. images						✓				
xView [62]	Object classification	30+ countries	Unknown	Sat. images						✓				
xBD (xView2) [41]	Disaster damage assessment	10 countries	2011-19	Sat. images						✓				
xView3 [4]	Illegal fishing detection	Oceans	Unknown	Sat. images									✓	
BigEarthNet [89]	Land cover classification	10 countries in Europe	2017-18	Sat. images										✓
ForestNet [50]	Deforestation drivers	Indonesia	2001-16	Environ. data & sat. images							✓			✓
iWildCam2020 / WILDS [13, 59]	Wildlife monitoring	12 countries	2013-15	Camera trap images										✓

remaining challenges in the domain (*e.g.*, learning from weak labels, sharing knowledge across the globe). Appendix D provides task-by-task comparisons of SUSTAINBENCH datasets with prior work.

3 SUSTAINBENCH Datasets and Tasks

In this section, we introduce the SUSTAINBENCH datasets and provide background on the SDGs that they help monitor. Seven SDGs are currently covered: No Poverty (SDG 1), Zero Hunger (SDG 2), Good Health and Well-being (SDG 3), Quality Education (SDG 4), Clean Water and Sanitation (SDG 6), Climate Action (SDG 13), and Life on Land (SDG 15). We describe how progress toward each goal is traditionally monitored, the gaps that currently exist in monitoring, and how certain indicators can be monitored using non-traditional datasets instead. Figure 1 summarizes the SDG, inputs, outputs, tasks, and original reference of each dataset, and Figures 2 and A1 visualize how many SDG indicators are covered by SUSTAINBENCH in each country. All of the datasets are easily downloaded via a Python package that integrates with the PyTorch ML framework [75].

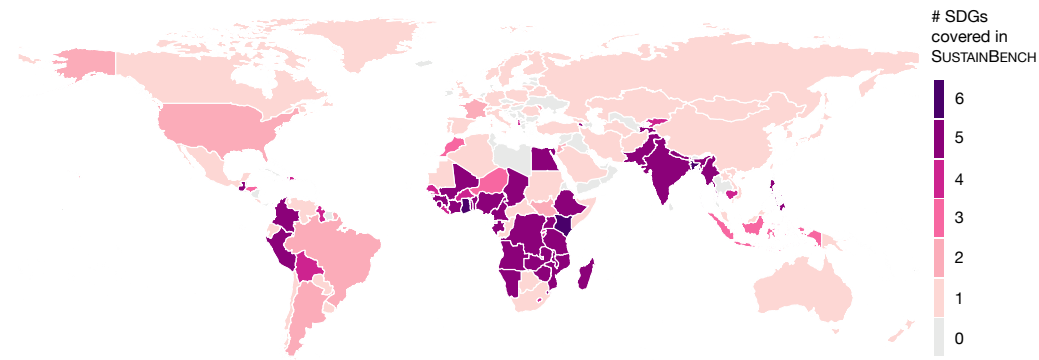


Figure 2: A map of how many SDGs are covered in SUSTAINBENCH for every country. SUSTAINBENCH has global coverage with an emphasis on low-income countries. In total, 119 countries have at least one task in SUSTAINBENCH.

3.1 No Poverty (SDG 1)

Despite decades of declining poverty rates, an estimated 8.4% of the global population remains in extreme poverty as of 2019, and progress has slowed in recent years [93]. But data on poverty remain surprisingly sparse, hampering efforts at monitoring local progress, targeting aid to those who need it, and evaluating the effectiveness of antipoverty programs [20]. In most African countries, for example, nationally representative consumption or asset wealth surveys, the key source of internationally comparable poverty measurements, are only available once every four years or less [109].

For SUSTAINBENCH, we processed survey data from two international household survey programs: Demographic and Health Surveys (DHS) [48] and the Living Standards Measurement Study (LSMS). Both constitute nationally representative household-level data on assets, housing conditions, and education levels, among other attributes. Notably, only LSMS data form a panel—*i.e.*, the same households are surveyed over time, facilitating comparison over time. Using a principal components analysis (PCA) approach [31, 85], we summarize the survey data into a single scalar asset wealth index per “cluster,” which roughly corresponds to a village or local community. We refer to cluster-level wealth (or its absence) as “poverty”. Previous research has shown that widely-available imagery sources including satellite imagery [52, 109] and crowd-sourced street-level imagery [64] can be effective for predicting cluster-level asset wealth when used as inputs in deep learning models.

SUSTAINBENCH includes two regression tasks for poverty prediction at the cluster level, both using imagery inputs to estimate an asset wealth index. The first task (Section 3.1.1) predicts poverty over space, and the second task (Section 3.1.2) predicts poverty changes over time.

3.1.1 Poverty Prediction Over Space

The poverty prediction over space task involves predicting a cluster-level asset wealth index which represents the “static” asset wealth of a cluster at a given point in time. For this task, the labels and inputs are created in a similar manner as in [109], but with about $5\times$ as many examples.

Dataset Following techniques developed in previous works [52, 109], we assembled asset wealth data for 2,079,036 households living in 86,936 clusters across 48 countries, drawn from DHS surveys conducted between 1996 and 2019, computing a cluster-level asset wealth index as described above. We provide satellite and street-level imagery inputs, gathered and processed according to established procedures [109, 64]. The $255\times 255\times 8$ px satellite images have 7 multispectral bands from Landsat daytime satellites and 1 nightlights band from either the DMSP or VIIRS satellites. The images are rescaled to a resolution of 30m/px and are geographically centered around each surveyed cluster’s geocoordinates. Geocoordinates in the public survey data are “jittered” by up to 10km from the true locations to protect the privacy of surveyed households [19]. For each cluster location, we also retrieved up to 300 crowd-sourced, street-level imagery from Mapillary. We evaluate model performance using the squared Pearson correlation coefficient (r^2) between predicted and observed values of the asset wealth index on held-out test countries. Appendix D.1 has more dataset details.

3.1.2 Poverty Prediction Over Time

For predicting temporal changes in poverty, we construct a PCA-based index of changes in asset ownership using LSMS data. For this task, the labels and inputs provided are similar to [109], with small improvements in image and label quality.

Dataset We provide labels for 1,665 instances of cluster-level asset wealth change from 1,287 clusters in 5 African countries. We use the same satellite imagery sources from the previous poverty prediction task. In this task, however, for each cluster we provide images from the two points in time (before and after) used to compute the difference in asset ownership, instead of only from a single point in time. Because street-level images were only available for $\sim 1\%$ of clusters, we do not provide them for this task. We evaluate model performance using the squared Pearson correlation coefficient (r^2) on predictions and labels in held-out cluster locations. Appendix D.2 has more dataset details.

3.2 Zero Hunger (SDG 2)

The number of people who suffer from hunger has risen since 2015, with 690 million or 9% of the world’s population affected by chronic hunger [93]. At the same time, 40% of habitable land on Earth is already devoted to agricultural activities, making agriculture by far the largest human impact on

the natural landscape [5]. The second SDG is to “end hunger, achieve food security and improved nutrition, and promote sustainable agriculture.” In addition to ending hunger and malnutrition in all forms, the targets under SDG 2 include doubling the productivity of small-scale food producers and promoting sustainable food production [93]. While traditionally data on agricultural practices and farm productivity are obtained via farm surveys, such data are rare and often of low quality [20]. Satellite imagery offers the opportunity to monitor agriculture more cheaply and more accurately, by mapping cropland, crop types, crop yields, field boundaries, and agricultural practices like cover cropping and conservation tillage. We discuss the SUSTAINBENCH datasets for SDG 2 below.

3.2.1 Cropland mapping with weak labels

One indicator for SDG 2 is the proportion of agricultural area under productive and sustainable agriculture [93]. Existing state-of-the-art datasets on land cover [18, 35] are derived from satellite time series and include a cropland class. However, the maps are known to have large errors in regions of the world like Sub-Saharan Africa where ground labels are sparse [56]. Therefore, while mapping cropland is largely a solved problem in settings with ample labels, devising methods to efficiently generate georeferenced labels and accurately map cropland in low-resource regions remains an important and challenging research direction.

Dataset We release a dataset for performing weakly supervised cropland classification in the U.S. using data from [102], which has not been released previously. While densely segmented labels are time-consuming and infeasible to generate for a large region like Africa, pixel-level and image-level labels are easier to create. The inputs are image tiles taken by the Landsat satellites and composited over the 2017 growing season, and the labels are either binary {cropland, not cropland} at single pixels or { $\geq 50\%$ cropland, $< 50\%$ cropland} for the entire image. Labels are generated from a high-quality USDA dataset on land cover [69]. Train, validation, and test sets are split along geographic blocks, and we evaluate models by overall accuracy and F1-score. We also encourage the use of semi-supervised and active learning methods to relieve the labeling burden needed to map cropland.

3.2.2 Crop type mapping in Sub-Saharan Africa

Spatially disaggregated crop type maps are needed to assess agricultural diversity and estimate yields. In high-income countries across North America and Europe, crop type maps are produced annually by departments of agriculture using farm surveys and satellite imagery [69]. However, no such maps are regularly available for middle- and low-income countries. Mapping crop types in the Global South faces challenges of irregularly shaped fields, small fields, intercropping, sparse ground truth labels, and highly heterogeneous landscapes [83]. We release two crop type datasets in Sub-Saharan Africa and point the reader to additional datasets hosted by the Radiant Earth Foundation [77] (Table 1). We recommend that ML researchers use all available datasets to ensure model generalizability.

Dataset #1 We re-release the dataset from [83] in Ghana and South Sudan in a format more familiar to the ML community. The inputs are growing season time series of imagery from three satellites (Sentinel-1, Sentinel-2, and PlanetScope) in 2016 and 2017, and the outputs are semantic segmentation of crop types. Ghana samples are labeled for maize, groundnut, rice, and soybean, while South Sudan samples are labeled for maize, groundnut, rice, and sorghum. We use the same train, validation, and test sets as [83], which preserve relative percentages of crop types across the splits. We evaluate models using overall accuracy and macro F1-score.

Dataset #2 We release the dataset used in [58] and [54] to map crop types in three regions of Kenya. Since the timing of growth and spectral signature are two main ways to distinguish crop types, the inputs are annual time series from the Sentinel-2 multi-spectral satellite. The outputs are crop types (9 possible classes). There are a total of 39,762 pixels belonging to 5,746 fields. The training, validation, and test sets are split along region rather than by field in order to develop models that generalize across geography. Our evaluation metrics are overall accuracy and macro-F1 score.

3.2.3 Crop yield prediction in North and South America

In order to double the productivity (or yield) of smallholder farms, we first have to measure it, and accurate local-level yield measurements are exceedingly rare in most of the world. In SUSTAINBENCH, we release county-level yields collected from various government databases; these can still aid in forecasting production, evaluating agricultural policy, and assessing the effects of climate change.

Dataset Our dataset is based on the datasets used in [110] and [101]. We release county-level yields for 857 counties in the U.S., 135 in Argentina, and 32 in Brazil for the years 2005-16. The inputs are spectral band and temperature histograms over each county for the harvest season from the MODIS satellite. The ground truth labels are the regional soybean yield per harvest, in metric tonnes per cultivated hectare, retrieved from government data. See Appendix D.6 for more details. Models are evaluated using root mean squared error (RMSE) and R^2 of predictions with the ground truth. The imbalance of data by country motivates the use of transfer learning approaches.

3.2.4 Field delineation in France

Since agricultural practices are usually implemented on the level of an entire field, field boundaries can help reduce noise and improve performance when mapping crop types and yields. Furthermore, field boundaries are a prerequisite for today’s digital agriculture services that help farmers optimize yields and profits [98]. Statistics that can be derived from field delineation, such as the size and distribution of crop fields, have also been used to study productivity [21, 27], mechanization [61], and biodiversity [37]. Field boundary datasets are rare and only sparsely labeled in low-income regions, so we release a large dataset from France to aid in model development.

Dataset We re-release the dataset introduced in Aung et al. 9. The dataset consists of Sentinel-2 satellite imagery in France over 3 time ranges: January-March, April-June, and July-September in 2017. The image has resolution 224×224 corresponding to a $2.24\text{km} \times 2.24\text{km}$ area on the ground. Each satellite image comes along with the corresponding binary masks of boundaries and areas of farm parcels. The dataset consists of a total of 1966 samples. We use a different data split from [9] to remove overlapping between the train, validation and test split. Following [9], we use the Dice score between the ground truth boundaries and predicted boundaries as the performance metric.

3.3 Good Health and Well-being (SDG 3)

Despite significant progress on improving global health outcomes (*e.g.*, halving child mortality rates since 2000 [93]), the lack of local-level measurements in many developing countries continues to constrain the monitoring, targeting, and evaluation of health interventions. We examine two health indicators: female body mass index (BMI), a key input to understanding both food insecurity and obesity; and child mortality rate (deaths under age 5), an official SDG 3 indicator considered to be a summary measure of a society’s health. Previous works have demonstrated using satellite imagery [67] or street-level Mapillary imagery inputs [64] for predicting BMI. While we are unaware of any prior works using such imagery inputs for predicting child mortality rates, “there is evidence that child mortality is connected to environmental factors such as housing quality, slum-like conditions, and neighborhood levels of vegetation” [51], which are certainly observable in imagery.

Dataset We provide cluster-level average labels for women’s BMI and child mortality rates compiled from DHS surveys. There are 94,866 cluster-level BMI labels computed from 1,781,403 women of childbearing age (15-49), excluding pregnant women. There are 105,582 cluster-level labels for child mortality rates computed from 1,936,904 children under age 5. As in the poverty prediction over space task (Section 3.1.1), the inputs for predicting the health labels are satellite and street-level imagery, and models are evaluated using the r^2 metric on labels from held-out test countries.

3.4 Quality Education (SDG 4)

SDG 4 includes targets that by 2030, all children and adults “complete free, equitable and quality primary and secondary education”. Increasing educational attainment (measured by years of schooling completed) is known to increase wealth and social mobility, and higher educational attainment in women is strongly associated with improved child nutrition and decreased child mortality [40]. Previous works have demonstrated the ability of deep learning methods to predict educational attainment from both satellite images [112] and street-level images [36, 64].

Dataset We provide cluster-level average years of educational attainment by women of reproductive age (15-49) compiled from same DHS surveys used for creating the asset wealth labels in the poverty prediction task. The 122,435 cluster-level labels were computed from 3,013,286 women across 56 countries. As in the poverty prediction over space task (Section 3.1.1), the inputs for predicting women educational attainment are satellite and street-level imagery, and models are evaluated using the r^2 metric on labels from held-out test countries.

3.5 Clean Water and Sanitation (SDG 6)

Clean water and sanitation are fundamental to human health, but as of 2020, two billion people globally do not have access to safe drinking water, and 2.3 billion lack a basic hand-washing facility with soap and water [84]. Access to improved sanitation and clean water is known to be associated with lower rates of child mortality [65, 33].

Dataset We provide cluster-level average years of a water quality index and sanitation index compiled from same DHS surveys used for creating the asset wealth labels in the poverty prediction task. The 87,938 (water index) and 89,271 (sanitation index) cluster-level labels were computed from 2,105,026 (water index) and 2,143,329 (sanitation index) households across 49 countries. As in the poverty prediction over space task (Section 3.1.1), the inputs for predicting the water quality and sanitation indices are satellite and street-level imagery, and models are evaluated using the r^2 metric on labels from held-out test countries. Since SUSTAINBENCH includes labels for child mortality in many of the same clusters with sanitation index labels, we encourage researchers to take advantage of the known associations between these variables.

3.6 Climate Action (SDG 13)

SDG 13 aims at combating climate change and its disruptive impacts on national economies and local livelihoods [68]. Monitoring emissions and environmental regulatory compliance are key steps toward SDG 13.

3.6.1 Brick kiln mapping

Brick manufacturing is a major source of carbon emissions and air pollution in South Asia, with an industry largely comprised of small-scale, informal producers. Identifying brick kilns from satellite imagery is a scalable method to improve compliance with environmental regulations and measure their impact on nearby populations. A recent study [63] trained a CNN to detect kilns and hand-validated the predictions, providing ground truth kiln locations in Bangladesh from October 2018 to May 2019.

Dataset The high-resolution satellite imagery used in [63] could not be shared publicly because they were proprietary. Hence, we provide a lower resolution alternative—Sentinel-2 imagery, which is available through Google Earth Engine [39]. We retrieved $64 \times 64 \times 13$ tiles at 10m/pixel resolution from the same time period and labeled each image as not containing a brick kiln (class 0) or containing a brick kiln (class 1) based on the ground truth locations in [63]. There were 6,329 positive examples out of 374,000 examples total; we sampled 25% of the negative examples and removed null values, resulting in 67,284 negative examples. More details can be found in Appendix D.8.

3.7 Life on Land (SDG 15)

Human activity has altered over 75% of the earth’s surface, reducing forest cover, degrading once-fertile land, and threatening an estimated 1 million animal and plant species with extinction [93]. Our understanding of land cover—*i.e.*, the physical material on the surface of the earth—and its changes is not uniform across the globe. Existing state-of-the-art land cover maps [18] are significantly more accurate in high-income regions than low-income ones, as the latter have few ground truth labels [56]. The following two datasets seek to reduce this gap via representation learning and transfer learning.

3.7.1 Representation learning for land cover classification

One approach to increase the performance of land cover classification in regions with few labels is to use unsupervised or self-supervised learning to improve satellite/aerial image representations, so that downstream tasks require fewer labels to perform well.

Dataset We release the high-resolution aerial imagery dataset from [53], which spans a 2500km² (12 billion pixel) area of Central Valley, CA in the U.S. The output is image-level land cover (66 classes), where labels are generated from a high-quality USDA dataset [69]. The region is divided in geographically-continuous blocks into train, validation, and test sets. The user may use the training imagery in any way to learn representations, and we provide a test set of up to 200,000 tiles (100×100px) for evaluation. The evaluation metrics are overall accuracy and macro F1-score.

Table 2: Benchmark performance on 15 tasks across 7 SDGs. See details in Appendix E. For the Model Type column, kNN = k-nearest neighbors, GP = Gaussian process. An asterisk (*) indicates a result on a similar dataset, but not the exact SUSTAINBENCH test set.

SDG	Task	Countries	Metric	Benchmark Value	Model Type	Ref
No Poverty	Poverty prediction over space	48 countries	r^2	0.63	kNN	[109]
	Poverty prediction over time	5 African countries	r^2	0.35*	ResNet-18	[109]
Zero Hunger	Weakly supervised cropland classification	United States	F1 score	0.88 (pixel label) 0.80 (image label)	U-Net	[102]
	Crop type classification	Ghana, South Sudan Kenya	Macro F1	0.57, 0.70	LSTM	[83]
			Macro F1	0.30	Random forest	[58]
	Crop yield prediction	United States Argentina, Brazil	RMSE	0.37 t/ha	CNN+GP	[110]
				0.62 t/ha, 0.42 t/ha	LSTM	[101]
Field delineation	France	Dice score	0.61 0.87	U-Net FracTAL Res-UNet	[9] [99]	
Good Health & Well-Being	Child mortality rate	56 countries	r^2	0.01	kNN	-
	Women BMI	53 countries	r^2	0.42	kNN	-
Quality Education	Women education	53 countries	r^2	0.26	kNN	-
	Water index	49 countries	r^2	0.40	kNN	-
Clean Water and Sanitation	Sanitation index	49 countries	r^2	0.36	kNN	-
	Brick kiln detection	Bangladesh	Accuracy	0.94*	ResNet-50	[63]
Life on Land	Representation learning for land cover	United States	Accuracy	0.55 ($n = 1,000$) 0.58 ($n = 10,000$)	Tile2Vec with ResNet-50	[53]
	Out-of-domain land cover classification	Global	Kappa	0.32 (1-shot, 2-way)	MAML with shallow 1D CNN	[104]

3.7.2 Out-of-domain land cover classification

A second strategy for increasing performance in label-scarce regions is to transfer knowledge learned from classifying land cover in high-income regions to low-income ones.

Dataset We release the global dataset of satellite time series from [104]. The dataset samples 692 regions of size $10\text{km} \times 10\text{km}$ around the globe; for each region, 500 latitude/longitude coordinates are sampled. The input is time series from the MODIS satellite over the course of a year, and the output is land cover type (17 possible classes). Users have the option of splitting regions into train, validation, and test sets at random or by continent. The evaluation metrics are overall accuracy, F1-score, and kappa score. The results from [104] are reported with all regions from Africa as the test set, but the user can choose to hold out other continents, for which the label quality will be higher.

4 Results for Baseline Models

SUSTAINBENCH provides a benchmark and public leaderboard website for the datasets described in Section 3. Each dataset has standard train-test splits with well-defined performance metrics detailed in Appendix E. We also welcome community submissions using additional data sources beyond what is provided in SUSTAINBENCH, such as for pre-training or regularization. Table 2 summarizes the baseline models and results. Code to reproduce our baseline models is available on GitHub⁴.

Here, we highlight some main takeaways from our baseline models. First, there is significant room for improvement for models that can take advantage of multi-modal inputs. Specifically, our baseline model for the DHS survey-based tasks only uses the satellite imagery inputs, and its poor performance on predicting child mortality and women educational attainment demonstrates the need to leverage additional data sources, such as the street-level imagery we provide. Second, ML model development can lead to significant gains in performance for SDG-related tasks. While the original paper that compiled SUSTAINBENCH’s field delineation dataset achieved a Dice score of 0.61 with a standard U-Net [9], we applied a new attention-based CNN developed specifically for field delineation [99] and achieved a 0.87 Dice score. For more task-specific discussions, please see Appendix E.

5 Impact, Limitations, and Future Work

This paper introduces SUSTAINBENCH, which, to the best of our knowledge, is the largest compilation to date of datasets and benchmarks for monitoring the SDGs with machine learning (ML). The SDGs

⁴<https://github.com/sustainlab-group/sustainbench/>

are arguably the most urgent challenges the world faces today, and it is important that the ML community contribute to solving these global issues. As progress towards SDGs is often hindered by a lack of ground survey data especially in low-income countries, ML algorithms designed for monitoring SDGs are important for leveraging non-traditional data sources that are cheap, globally available, and frequently-updated to fill in data gaps. ML-based estimates provide policymakers from governments and aid organizations with more frequent and comprehensive insights [109, 20, 52].

The tasks defined in SUSTAINBENCH can directly translate into real-world impact. For example, during the COVID-19 pandemic, the government of Togo collaborated with researchers to use satellite imagery, phone data, and ML to map poverty [14] and cropland [56] in order to target cash payments to the jobless. Recent work in Uganda demonstrates how ML-based poverty maps can be used to measure the effectiveness of large-scale infrastructure investments [78]. ML-based analyses of satellite images in Kenya (using the labels described in Section 3.2.2) were recently used to identify soil nitrogen deficiency as the limiting factor in maize yields, thereby facilitating targeted agriculture intervention [54]. And as a last example, the development of a new attention-based neural network architecture enabled the delineation of 1.7 million fields in Australia from satellite imagery [99]. These field boundaries have been productized and facilitate the adoption of digital agriculture, which can improve yields while minimizing environmental pollution [24].

Although ML approaches have demonstrated value on a variety of tasks related to SDGs [109, 20, 64, 53, 52, 101, 103], the “big data approach” has its limits. ML models may not completely replace ground surveys. Imperfect predictions from ML models may introduce biases that propagate through downstream policy decisions, leading to negative societal impacts. The use of survey data, high resolution remote sensing images, and street-level images may also raise privacy concerns, despite efforts to protect individual privacy. We refer the reader to Appendix F for a detailed treatment of ethical concerns in SUSTAINBENCH, including mitigation strategies we implemented. Despite these limitations, ML applications have the greatest potential for positive impact in low-income countries, where gaps in monitoring SDGs are widest due to the constant lack of survey data.

While SUSTAINBENCH is the largest SDG-focused ML dataset and benchmark to date, it is by no means complete. Field surveys are extremely costly, and labeling images for model training requires significant manual effort by experts, limiting the amount of data released in SUSTAINBENCH to quantities smaller than those of many canonical ML datasets (*e.g.*, ImageNet). In addition, many SDGs and indicators are not included in the current version. Such SDG indicators can be placed into 3 categories. First, several tasks can be included in future versions of SUSTAINBENCH by drawing on existing data. For example, measures of gender equality (SDG 5) and access to affordable and clean energy (SDG 7) already exist in the surveys used to create labels for SUSTAINBENCH tasks but will require additional processing before releasing. Recent works have also pioneered deep learning methods for identifying illegal fishing from satellite images [74] (SDG 14) and monitoring biodiversity from camera traps [13] (SDG 15). Table 1 includes a few relevant datasets from this first category. Second, some SDG indicators require additional research to discover non-traditional data modalities that can be used to monitor them. Finally, not all SDGs are measurable using ML or need improved measurement capabilities from ML models. For example, international cooperation (SDG 17) is perhaps best measured by domestic and international policies and agreements.

For the ML community, SUSTAINBENCH also provides opportunities to test state-of-the-art ML models on real-world data and develop novel algorithms. For example, the tasks based on DHS household survey data share the same inputs and thus facilitate multi-task training. In particular, we encourage researchers to take advantage of the known strong associations between asset wealth, child mortality, women’s education, and sanitation labels [33, 40]. The combination of satellite and street-level imagery for these tasks also enables multi-modal representation learning. On the other hand, the land cover classification and cropland mapping tasks provide new real-world datasets for evaluating and developing self-supervised, weakly supervised, unsupervised, and meta-learning algorithms. We welcome exploration of methods beyond our provided baseline models.

Ultimately, we hope SUSTAINBENCH will lower the barrier to entry for the ML community to contribute toward monitoring SDGs and highlight challenges for ML researchers to address. In the long run, we plan to continue expanding datasets and benchmarks as new data sources become available. We believe that standardized datasets and benchmarks like those in SUSTAINBENCH are imperative to both novel method development and real-world impact.

Acknowledgments

The authors would like to thank everyone from the Stanford Sustainability and AI Lab for constructive feedback and discussion; the Mapillary team for technical support on the dataset; Rose Rustowicz for helping compile the crop type mapping dataset in Ghana and South Sudan; Anna X. Wang and Jiaxuan You for their help in making the crop yield dataset; and Han Lin Aung and Burak Uz kent for permission to release the field delineation dataset.

This work was supported by NSF awards (#1651565, #1522054), the Stanford Institute for Human-Centered AI (HAI), the Stanford King Center, the United States Agency for International Development (USAID), a Sloan Research Fellowship, and the Global Innovation Fund.

References

- [1] Crop yield forecasting, Nov 2012. URL <https://ec.europa.eu/jrc/en/research-topic/crop-yield-forecasting>.
- [2] Transforming our World: The 2030 Agenda for Sustainable Development, Sep 2015. URL <https://sustainabledevelopment.un.org/post2015/transformingourworld/publication>.
- [3] Blurring images. <https://help.mapillary.com/hc/en-us/articles/115001663705-Blurring-images>, 2021.
- [4] xView3: Dark Vessels, 2021. URL <https://iuu.xview.us/>.
- [5] Food and Agriculture Statistics, 2021. URL <http://www.fao.org/food-agriculture-statistics/en/>.
- [6] E. Aiken, S. Bellue, D. Karlan, C. R. Udry, and J. Blumenstock. Machine Learning and Mobile Phone Data Can Improve the Targeting of Humanitarian Assistance. Working Paper 29070, National Bureau of Economic Research, Jul 2021. URL <https://www.nber.org/papers/w29070>.
- [7] S. Alkire, J. M. Roche, P. Ballon, J. Foster, M. E. Santos, and S. Seth. *Multidimensional Poverty Measurement and Analysis*. Oxford University Press, New York, NY, USA, 1 edition, 2015. ISBN 978-0-19-968949-1.
- [8] Argentina Subsecretaría de Agricultura. Estimaciones agrícolas. URL <http://datosestimaciones.magyp.gob.ar/reportes.php?reporte=Estimaciones>.
- [9] H. L. Aung, B. Uz kent, M. Burke, D. Lobell, and S. Ermon. Farm parcel delineation using spatio-temporal convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 76–77, 2020.
- [10] G. Azzari and D. B. Lobell. Landsat-based classification in the cloud: An opportunity for a paradigm shift in land cover monitoring. *Remote Sensing of Environment*, pages 1–11, May 2017.
- [11] G. Azzari, M. Jain, and D. B. Lobell. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. *Remote Sensing of Environment*, 202:129–141, 2017.
- [12] B. Babenko, J. Hersh, D. Newhouse, A. Ramakrishnan, T. Swartz, and W. Bank. Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images, With an Application in Mexico. In *NIPS 2017 Workshop on Machine Learning for the Developing World*, 2017. URL <https://arxiv.org/abs/1711.06323>.
- [13] S. Beery, E. Cole, and A. Gjoka. The iWildCam 2020 Competition Dataset. *arXiv preprint arXiv:2004.10340*, 2020.
- [14] J. Blumenstock. Machine learning can help get COVID-19 aid to those who need it most. *Nature*, May 2020. doi: 10.1038/d41586-020-01393-7. URL <https://www.nature.com/articles/d41586-020-01393-7>.
- [15] J. Blumenstock, G. Cadamuro, and R. On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.
- [16] D. K. Bolton and M. A. Friedl. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, 173:74–84, 2013. ISSN 0168-1923. doi: 10.1016/j.agrformet.2013.01.007. URL <https://www.sciencedirect.com/science/article/pii/S0168192313000129>.

- [17] Brasil Sistema IBGE de Recuperacao Automatica, Instituto Brasileiro de Geografia e Estatistica. Producao agricola municipal: producao das lavouras temporárias. URL <https://sidra.ibge.gov.br/tabela/1612>.
- [18] M. Buchhorn, M. Lesiv, N.-E. Tsendbazar, M. Herold, L. Bertels, and B. Smets. Copernicus Global Land Cover Layers—Collection 2. *Remote Sensing*, 12(6), 2020. ISSN 2072-4292. doi: 10.3390/rs12061044. URL <https://www.mdpi.com/2072-4292/12/6/1044>.
- [19] C. R. Burgert, J. Colston, T. Roy, and B. Zachary. Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys. 2013. URL <http://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf>.
- [20] M. Burke, A. Driscoll, D. B. Lobell, and S. Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021. doi: 10.1126/science.abe8628. URL <https://www.science.org/doi/abs/10.1126/science.abe8628>.
- [21] M. R. Carter. Identification of the inverse relationship between farm size and productivity: An empirical analysis of peasant agricultural production. *Oxford Economic Papers*, 36(1):131–145, 1984. ISSN 00307653, 14643812. URL <http://www.jstor.org/stable/2662637>.
- [22] R. Chew, J. Rineer, R. Beach, M. O’Neil, N. Ujeneza, D. Lapidus, T. Miano, M. Hegarty-Craver, J. Polly, and D. S. Temple. Deep Neural Networks and Transfer Learning for Food Crop Identification in UAV Images. *Drones*, 4(1), 2020.
- [23] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.
- [24] CSIRO. ePaddocks Australian Paddock Boundaries. URL <https://acds.csiro.au/epaddock-australian-paddock-boundaries>.
- [25] J. A. de Bruijn, H. de Moel, B. Jongman, M. C. de Ruiter, J. Wagemaker, and J. C. J. H. Aerts. A global database of historic and real-time flood events based on social media. *Scientific Data*, 6(1):311, 2019.
- [26] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raska. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In *CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–17209, Jun 2018. doi: 10.1109/CVPRW.2018.00031.
- [27] S. Desiere and D. Jolliffe. Land productivity and plot size: Is measurement error driving the inverse relationship? *Journal of Development Economics*, 130:84–98, 2018.
- [28] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=HkpbH9lx>.
- [29] C. D. Elvidge, K. Baugh, M. Zhizhin, F. C. Hsu, and T. Ghosh. VIIRS night-time lights. *International Journal of Remote Sensing*, 38(21):5860–5879, June 2017. ISSN 0143-1161. doi: 10.1080/01431161.2017.1342050. URL <https://www.tandfonline.com/doi/10.1080/01431161.2017.1342050>.
- [30] R. Engstrom, J. S. Hersh, and D. L. Newhouse. Poverty from space: using high-resolution satellite imagery for estimating economic well-being. Technical report, World Bank Group, Washington, D.C., 2017. URL <http://documents.worldbank.org/curated/en/610771513691888412/Poverty-from-space-using-high-resolution-satellite-imagery-for-estimating-economic-well-being>.
- [31] D. Filmer and L. H. Pritchett. Estimating Wealth Effects Without Expenditure Data—Or Tears: An Application To Educational Enrollments In States Of India. *Demography*, 38(1):115–132, Feb 2001. ISSN 1533-7790. doi: 10.1353/dem.2001.0003. URL <https://doi.org/10.1353/dem.2001.0003>.
- [32] D. Filmer and K. Scott. Assessing Asset Indices. *Demography*, 49(1):359–392, Feb 2012. ISSN 1533-7790. doi: 10.1007/s13524-011-0077-5. URL <https://doi.org/10.1007/s13524-011-0077-5>.
- [33] G. Fink, I. Günther, and K. Hill. The effect of water and sanitation on child health: evidence from the demographic and health surveys 1986–2007. *International Journal of Epidemiology*, 40(5):1196–1204, Oct 2011. ISSN 0300-5771. doi: 10.1093/ije/dyr102. URL <https://doi.org/10.1093/ije/dyr102>.
- [34] M. Friedl and D. Sulla-Menashe. MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006. 2019. doi: 10.5067/MODIS/MCD12Q1.006. URL <https://lpdaac.usgs.gov/products/mcd12q1v006/>.

- [35] M. Friedl, D. McIver, J. Hodges, X. Zhang, D. Muchoney, A. Strahler, C. Woodcock, S. Gopal, A. Schneider, A. Cooper, A. Baccini, F. Gao, and C. Schaaf. Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, 83(1):287–302, 2002.
- [36] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, Dec 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1700035114. URL <https://www.pnas.org/content/114/50/13108>.
- [37] F. Geiger, J. Bengtsson, F. Berendse, W. W. Weisser, M. Emmerson, M. B. Morales, P. Ceryngier, J. Liira, T. Tschardtke, C. Winqvist, S. Eggers, R. Bommarco, T. Pärt, V. Bretagnolle, M. Plantegenest, L. W. Clement, C. Dennis, C. Palmer, J. J. Oñate, I. Guerrero, V. Hawro, T. Aavik, C. Thies, A. Flohre, S. Hänke, C. Fischer, P. W. Goedhart, and P. Inchausti. Persistent negative effects of pesticides on biodiversity and biological control potential on European farmland. *Basic and Applied Ecology*, 11(2):97–105, 2010.
- [38] U. Gentilini, S. Khosla, and M. Almenfi. Cash in the City: Emerging Lessons from Implementing Cash Transfers in Urban Africa. Technical report, World Bank, Washington, D.C., USA, Jan 2021. URL <https://openknowledge.worldbank.org/handle/10986/35003>.
- [39] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017. doi: 10.1016/j.rse.2017.06.031. URL <https://doi.org/10.1016/j.rse.2017.06.031>.
- [40] N. Graetz, J. Friedman, A. Osgood-Zimmerman, R. Burstein, M. H. Biehl, C. Shields, J. F. Mosser, D. C. Casey, A. Deshpande, L. Earl, R. C. Reiner, S. E. Ray, N. Fullman, A. J. Levine, R. W. Stubbs, B. K. Mayala, J. Longbottom, A. J. Browne, S. Bhatt, D. J. Weiss, P. W. Gething, A. H. Mokdad, S. S. Lim, C. J. L. Murray, E. Gakidou, and S. I. Hay. Mapping local variation in educational attainment across Africa. *Nature*, 555(7694), Mar 2018. ISSN 1476-4687. doi: 10.1038/nature25761. URL <http://www.nature.com/articles/nature25761>.
- [41] R. Gupta, R. Hosfelt, S. Sajeev, N. Patel, B. Goodman, J. Doshi, E. Heim, H. Choset, and M. Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019.
- [42] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160): 850–853, 2013.
- [43] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL <https://ieeexplore.ieee.org/document/7780459>.
- [44] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [45] A. Head, M. Manguin, N. Tran, and J. E. Blumenstock. Can Human Development be Measured with Satellite Imagery? In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, pages 1–11, Lahore, Pakistan, Nov 2017. ACM. ISBN 978-1-4503-5277-2. doi: 10.1145/3136560.3136576. URL <http://dl.acm.org/citation.cfm?doid=3136560.3136576>.
- [46] F.-C. Hsu, K. Baugh, T. Ghosh, M. Zhizhin, and C. Elvidge. DMSP-OLS Radiance Calibrated Nighttime Lights Time Series with Intercalibration. *Remote Sensing*, 7(2):1855–1876, Feb 2015. doi: 10.3390/rs70201855. URL <http://www.mdpi.com/2072-4292/7/2/1855>.
- [47] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.
- [48] ICF. Demographic and Health Surveys (various), 1996–2019. Funded by USAID.
- [49] J. Inglada, M. Arias, B. Tardy, O. Hagolle, S. Valero, D. Morin, G. Dedieu, G. Sepulcre, S. Bontemps, P. Defourny, and B. Koetz. Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery. *Remote Sensing*, 7(9):12356–12379, 2015.

- [50] J. Irvin, H. Sheng, N. Ramachandran, S. Johnson-Yu, S. Zhou, K. Story, R. Rustowicz, C. Elsworth, K. Austin, and A. Y. Ng. ForestNet: Classifying Drivers of Deforestation in Indonesia using Deep Learning on Satellite Imagery. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*, Dec 2020. URL <https://www.climatechange.ai/papers/neurips2020/22>.
- [51] M. M. Jankowska, M. Benza, and J. R. Weeks. Estimating spatial inequalities of urban child mortality. *Demographic research*, 28:33–62, Jan 2013. ISSN 1435-9871. doi: 10.4054/DemRes.2013.28.2. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3903295/>.
- [52] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–4, Aug 2016. doi: 10.1126/science.aaf7894. URL <https://science.sciencemag.org/content/353/6301/790>.
- [53] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon. Tile2Vec: Unsupervised Representation Learning for Spatially Distributed Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3967–3974, Jul 2019.
- [54] Z. Jin, G. Azzari, C. You, S. Di Tommaso, S. Aston, M. Burke, and D. B. Lobell. Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sensing of Environment*, 228: 115–128, 2019.
- [55] H. Kerner, C. Nakalembe, and I. Becker-Reshef. Field-Level Crop Type Classification with k Nearest Neighbors: A Baseline for a New Kenya Smallholder Dataset, 2020.
- [56] H. Kerner, G. Tseng, I. Becker-Reshef, C. Nakalembe, B. Barker, B. Munshell, M. Paliyam, and M. Hosseini. Rapid Response Crop Maps in Data Sparse Regions. In *KDD '20: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) Humanitarian Mapping Workshop*. ACM, 8 2020. URL <https://arxiv.org/abs/2006.16866>.
- [57] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [58] D. M. Kluger, S. Wang, and D. B. Lobell. Two shifts for crop mapping: Leveraging aggregate crop statistics to improve satellite-based maps in new regions. *Remote Sensing of Environment*, 262:112488, 2021.
- [59] P. W. Koh, S. Sagawa, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, et al. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.
- [60] A. Krizhevsky. Learning multiple layers of features from tiny images. Master’s thesis, University of Toronto, Apr. 2009.
- [61] T. Kuemmerle, K. Erb, P. Meyfroidt, D. Müller, P. H. Verburg, S. Estel, H. Haberl, P. Hostert, M. R. Jepsen, T. Kastner, C. Levers, M. Lindner, C. Plutzer, P. J. Verkerk, E. H. van der Zanden, and A. Reenberg. Challenges and opportunities in mapping land use intensity globally. *Current Opinion in Environmental Sustainability*, 5(5):484–493, 2013.
- [62] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord. xView: Objects in Context in Overhead Imagery. *arXiv:1802.07856 [cs]*, Feb 2018. URL <http://arxiv.org/abs/1802.07856>.
- [63] J. Lee, N. R. Brooks, F. Tajwar, M. Burke, S. Ermon, D. B. Lobell, D. Biswas, and S. P. Luby. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences*, 118(17), 2021. ISSN 0027-8424. doi: 10.1073/pnas.2018863118. URL <https://www.pnas.org/content/118/17/e2018863118>.
- [64] J. Lee, D. Grosz, B. Uz Kent, S. Zeng, M. Burke, D. Lobell, and S. Ermon. Predicting Livelihood Indicators from Community-Generated Street-Level Imagery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):268–276, May 2021. ISSN 2374-3468. URL <https://ojs.aaai.org/index.php/AAAI/article/view/16101>.
- [65] Local Burden of Disease WaSH Collaborators. Mapping geographical inequalities in access to drinking water and sanitation facilities in low-income and middle-income countries, 2000–17. *The Lancet Global Health*, 8(9):e1162–e1185, Sep 2020. ISSN 2214-109X. doi: 10.1016/S2214-109X(20)30278-3. URL [https://www.thelancet.com/journals/langlo/article/PIIS2214-109X\(20\)30278-3/fulltext](https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(20)30278-3/fulltext).

- [66] C. Lowe, A. McCord, and R. Beazley. National cash transfer responses to Covid-19: operational lessons learned for social protection system-strengthening and future shocks. Technical Report Working Paper 610, Overseas Development Institute, June 2021. URL <https://odi.org/en/publications/national-cash-transfer-responses-to-covid-19-operational-lessons-learned-for-social-protection-system-strengthening-and-future-shocks/>.
- [67] A. Maharana and E. O. Nsoesie. Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity. *JAMA Network Open*, 1(4):e181535, Aug 2018. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2018.1535. URL <https://doi.org/10.1001/jamanetworkopen.2018.1535>.
- [68] Martin. Climate Change, Aug 2021. URL <https://www.un.org/sustainabledevelopment/climate-change/>.
- [69] National Agricultural Statistics Service. USDA National Agricultural Statistics Service Cropland Data Layer. Published crop-specific data layer [Online], 2018. URL <https://nassgeodata.gmu.edu/CropScape/>.
- [70] U. Nations. Prototype Global Sustainable Development Report. Technical report, United Nations Department of Economic and Social Affairs, Division for Sustainable Development, 2014.
- [71] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kotschieder. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In *International Conference on Computer Vision (ICCV)*, 2017. URL <https://www.mapillary.com/dataset/vistas>.
- [72] Neuhold, Gerhard. Accurate Privacy Blurring at Scale, 2018. URL <https://blog.mapillary.com/update/2018/04/19/accurate-privacy-blurring-at-scale.html>.
- [73] A. M. Noor, V. A. Alegana, P. W. Gething, A. J. Tatem, and R. W. Snow. Using remotely sensed night-time light as a proxy for poverty in Africa. *Population health metrics*, 6:5, Oct. 2008. ISSN 1478-7954. doi: 10.1186/1478-7954-6-5. URL <http://www.ncbi.nlm.nih.gov/pubmed/18939972>.
- [74] J. Park, J. Lee, K. Seto, T. Hochberg, B. A. Wong, N. A. Miller, K. Takasaki, H. Kubota, Y. Oozeki, S. Doshi, M. Midzik, Q. Hanich, B. Sullivan, P. Woods, and D. A. Kroodsma. Illuminating dark fishing fleets in North Korea. *Science Advances*, 6(30), 2020. doi: 10.1126/sciadv.abb1197. URL <https://advances.sciencemag.org/content/6/30/eabb1197>.
- [75] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.
- [76] N. A. Quarmby, M. Milnes, T. L. Hindle, and N. Silleos. The use of multi-temporal NDVI measurements from AVHRR data for crop yield estimation and prediction. *International Journal of Remote Sensing*, 14(2):199–210, 1993. doi: 10.1080/01431169308904332. URL <https://doi.org/10.1080/01431169308904332>.
- [77] Radiant Earth Foundation. Machine Learning for Earth Observation, 2021. URL <https://www.radiant.earth/mlhub/>.
- [78] N. Ratledge, G. Cadamuro, B. De la Cuesta, M. Stigler, and M. Burke. Using satellite imagery and machine learning to estimate the livelihood impact of electricity access. Technical report, National Bureau of Economic Research, 2021.
- [79] R. Remelgado, S. Zaitov, S. Kenjabaev, G. Stulina, M. Sultanov, M. Ibrakhimov, M. Akhmedov, V. Dukhovny, and C. Conrad. A crop type dataset for consistent land cover classification in Central Asia. *Scientific Data*, 7(1):250, 2020.
- [80] E. Rolf, J. Proctor, T. Carleton, I. Bolliger, V. Shankar, M. Ishihara, B. Recht, and S. Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1):4392, 2021.
- [81] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- [82] M. Russwurm, S. Wang, M. Korner, and D. Lobell. Meta-Learning for Few-Shot Land Cover Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

- [83] R. Rustowicz, R. Cheong, L. Wang, S. Ermon, M. Burke, and D. Lobell. Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [84] J. Sachs, C. Kroll, G. Lafortune, G. Fuller, and F. Woelm. *Sustainable Development Report 2021*. Cambridge University Press, 2021.
- [85] D. E. Sahn and D. Stifel. Exploring Alternative Measures of Welfare in the Absence of Expenditure Data. *Review of Income and Wealth*, 49(4):463–489, 2003. ISSN 1475-4991. doi: 10.1111/j.0034-6586.2003.00100.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0034-6586.2003.00100.x>.
- [86] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu. SEN12MS—A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion. *arXiv preprint arXiv:1906.07789*, June 2019.
- [87] E. Sheehan, C. Meng, M. Tan, B. UzKent, N. Jean, M. Burke, D. Lobell, and S. Ermon. Predicting Economic Development using Geolocated Wikipedia Articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, pages 2698–2706, New York, NY, USA, July 2019. Association for Computing Machinery. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330784. URL <https://doi.org/10.1145/3292500.3330784>.
- [88] Stanford Woods Institute for the Environment. A Better Brick: Solving an Airborne Health Threat, 2021. URL <https://woods.stanford.edu/research/funding-opportunities/environmental-venture-projects/brick-kiln-solutions>.
- [89] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019.
- [90] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai. County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model. *Sensors*, 19(20), 2019. ISSN 1424-8220. doi: 10.3390/s19204363. URL <https://www.mdpi.com/1424-8220/19/20/4363>.
- [91] D. Tedesco-Oliveira, R. Pereira da Silva, W. Maldonado, and C. Zerbatto. Convolutional neural networks in predicting cotton yield from images of commercial fields. *Computers and Electronics in Agriculture*, 171:105307, 2020. ISSN 0168-1699. doi: <https://doi.org/10.1016/j.compag.2020.105307>. URL <https://www.sciencedirect.com/science/article/pii/S0168169919319878>.
- [92] B. Tellman, J. A. Sullivan, C. Kuhn, A. J. Kettner, C. S. Doyle, G. R. Brakenridge, T. A. Erickson, and D. A. Slayback. Satellite imaging reveals increased proportion of population exposed to floods. *Nature*, 596(7870):80–86, 2021.
- [93] United Nations Department of Economic and Social Affairs. *The Sustainable Development Goals Report 2021*. The Sustainable Development Goals Report. United Nations, 2021 edition, 2021. ISBN 978-92-1-005608-3. doi: 10.18356/9789210056083. URL <https://www.un-ilibrary.org/content/books/9789210056083>.
- [94] United Nations Statistics Division. Tier Classification for Global SDG Indicators, 2021. URL <https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/>.
- [95] USDA. USDA National Agricultural Statistics Service. URL <https://www.nass.usda.gov/>.
- [96] A. Van Etten, D. Lindenbaum, and T. M. Bacastow. SpaceNet: A Remote Sensing Dataset and Challenge Series. *arXiv:1807.01232 [cs]*, July 2019. URL <http://arxiv.org/abs/1807.01232>.
- [97] E. Vermote. MOD09A1 MODIS/Terra Surface Reflectance 8-Day L3 Global 500m SIN Grid V006. 2015. doi: 10.5067/MODIS/MOD09A1.006. URL <https://lpdaac.usgs.gov/products/mod09a1v006/>.
- [98] F. Waldner and F. I. Diakogiannis. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. *Remote Sensing of Environment*, 245:111741, 2020.
- [99] F. Waldner, F. I. Diakogiannis, K. Batchelor, M. Ciccotosto-Camp, E. Cooper-Williams, C. Herrmann, G. Mata, and A. Toovey. Detect, consolidate, delineate: Scalable mapping of field boundaries using satellite images. *Remote Sensing*, 13(11), 2021.
- [100] Z. Wan, S. Hook, and G. Hulley. MYD11A2 MODIS/Aqua Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006. 2015. doi: 10.5067/MODIS/MYD11A2.006. URL <https://lpdaac.usgs.gov/products/myd11a2v006/>. Type: dataset.

- [101] A. X. Wang, C. Tran, N. Desai, D. Lobell, and S. Ermon. Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS '18*, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450358163. doi: 10.1145/3209811.3212707. URL <https://doi.org/10.1145/3209811.3212707>.
- [102] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12(2), 2020. doi: 10.3390/rs12020207.
- [103] S. Wang, S. Di Tommaso, J. Faulkner, T. Friedel, A. Kennepohl, R. Strey, and D. B. Lobell. Mapping Crop Types in Southeast India with Smartphone Crowdsourcing and Deep Learning. *Remote Sensing*, 12(18), 2020.
- [104] S. Wang, M. Rußwurm, M. Körner, and D. B. Lobell. Meta-learning for few-shot time series classification. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 7041–7044, 2020. doi: 10.1109/IGARSS39084.2020.9441016.
- [105] G. R. Watmough, C. L. J. Marcinko, C. Sullivan, K. Tschirhart, P. K. Mutuo, C. A. Palm, and J.-C. Svenning. Socioecologically informed use of remote sensing data to predict rural household poverty. *Proceedings of the National Academy of Sciences*, 116(4):1213–1218, Jan 2019. ISSN 0027-8424. doi: 10.1073/pnas.1812969116. URL <https://www.pnas.org/content/116/4/1213>.
- [106] J. Xiong, P. S. Thenkabail, M. K. Gumma, P. Teluguntla, J. Poehnelt, R. G. Congalton, K. Yadav, and D. Thau. Automated cropland mapping of continental Africa using Google Earth Engine cloud computing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 126:225–244, 2017.
- [107] L. Yan and D. Roy. Conterminous United States crop field size quantification from multi-temporal Landsat data. *Remote Sensing of Environment*, 172:67–86, 2016.
- [108] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450304283. doi: 10.1145/1869790.1869829. URL <https://doi.org/10.1145/1869790.1869829>.
- [109] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1), May 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16185-w. URL <https://www.nature.com/articles/s41467-020-16185-w>.
- [110] J. You, X. Li, M. Low, D. Lobell, and S. Ermon. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17*, page 4559–4565. AAAI Press, 2017. URL <https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14435>.
- [111] H. Zhao, S. Duan, J. Liu, L. Sun, and L. Reymondin. Evaluation of Five Deep Learning Models for Crop Type Mapping Using Sentinel-2 Time Series Images with Missing Information. *Remote Sensing*, 13(14), 2021.
- [112] S. Zhao, C. Yeh, and S. Ermon. A Framework for Sample Efficient Interval Estimation with Control Variates. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 4583–4592. PMLR, June 2020. URL <https://proceedings.mlr.press/v108/zhao20e.html>.