

## Appendix

### A Dataset Licenses

The Landsat, DMSP, NAIP, and VIIRS satellite images provided in SustainBench are in the public domain. PlanetScope imagery and Mapillary street-level imagery are provided under the CC BY-SA 4.0 license. Sentinel-2 imagery is provided under the Open Access compliant Creative Commons CC BY-SA 3.0 IGO license. Sentinel-1 imagery provides free access to imagery, including reproduction and distribution <sup>5</sup>. Likewise, MODIS imagery is free to reuse and redistribute <sup>6</sup>.

Our inclusion of labels derived from DHS survey data is within the DHS program Terms of Use<sup>7</sup> as the labels are aggregated to the cluster level and do not include any of the original “micro-level” data, and no individuals are identified.

Our inclusion of labels derived from LSMS survey data is within the LSMS access policy, as we do not redistribute any of the raw data files.

The Argentina crop yield labels are provided under the CC BY 2.5 AR license. United States crop yield labels are also free to access and reproduce <sup>8</sup>.

The brick kiln binary classification labels were manually hand-labeled by ourselves and our collaborators and therefore do not have any licensing restrictions.

SUSTAINBENCH itself is released under a CC BY-SA 4.0 license, which is compatible with all of the licenses for the datasets included.

### B Dataset Storage and Maintenance Plans

Our datasets are stored on Google Drive at the following link: <https://drive.google.com/drive/folders/1jyjK5sKGYegfHDjuVBSxCoj49TD830wL?usp=sharing>. Due to the large size of our dataset, we were unable to find any existing research data repository (*e.g.*, Zenodo, Dataverse) willing to accommodate our dataset.

The GitHub repo with code used to process the datasets and run our baseline models is located at <https://github.com/sustainlab-group/sustainbench/>.

The dataset will be maintained by the Stanford Sustainability and AI lab.

---

<sup>5</sup>[https://scihub.copernicus.eu/twiki/pub/SciHubWebPortal/TermsConditions/Sentinel\\_Data\\_Terms\\_and\\_Conditions.pdf](https://scihub.copernicus.eu/twiki/pub/SciHubWebPortal/TermsConditions/Sentinel_Data_Terms_and_Conditions.pdf)

<sup>6</sup><https://lpdaac.usgs.gov/data/data-citation-and-policies/>

<sup>7</sup><https://dhsprogram.com/data/terms-of-use.cfm>

<sup>8</sup>[https://www.nass.usda.gov/Data\\_and\\_Statistics/Citation\\_Request/index.php](https://www.nass.usda.gov/Data_and_Statistics/Citation_Request/index.php)

Table A1: The full list of 17 UN Sustainable Development Goals (SDGs), along with the number of targets and indicators divided by tier.

SDG #	Name	Description	# of Targets	# of Indicators		
				Tier I	Tier II	Tier I/II
1	No Poverty	End poverty in all its forms everywhere	7	5	8	0
2	Zero Hunger	End hunger, achieve food security and improved nutrition and promote sustainable agriculture	8	10	4	0
3	Good Health and Well-Being	Ensure healthy lives and promote well-being for all at all ages	13	25	3	0
4	Quality Education	Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all	10	5	6	1
5	Gender Equality	Achieve gender equality and empower all women and girls	9	4	10	0
6	Clean Water and Sanitation	Ensure availability and sustainable management of water and sanitation for all	8	7	4	0
7	Affordable and Clean Energy	Ensure access to affordable, reliable, sustainable and modern energy for all	5	6	0	0
8	Decent Work and Economic Growth	Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all	12	8	8	0
9	Industry, Innovation and Infrastructure	Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation	8	10	2	0
10	Reduced Inequalities	Reduce inequality within and among countries	10	8	6	0
11	Sustainable Cities and Communities	Make cities and human settlements inclusive, safe, resilient and sustainable	10	4	10	0
12	Responsible Consumption and Production	Ensure sustainable consumption and production patterns	11	5	8	0
13	Climate Action	Take urgent action to combat climate change and its impacts	5	2	6	0
14	Life below Water	Conserve and sustainably use the oceans, seas and marine resources for sustainable development	10	5	5	0
15	Life on Land	Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss	12	11	2	1
16	Peace, Justice and Strong Institutions	Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels	12	6	17	1
17	Partnerships for the Goals	Strengthen the means of implementation and revitalize the global partnership for sustainable development	19	15	8	1
<b>Total</b>			169	136	107	4

## C The 17 Sustainable Development Goals (SDGs)

Today, six years after the unveiling of the SDGs, many gaps still exist in monitoring progress. Official tracking of data availability is conducted by the UN Statistical Commission, which classifies each indicator into one of three tiers: indicator is well-defined and data are regularly produced by at least 50% of countries (Tier I), indicator is well-defined but data are not regularly produced by countries (Tier II), and the indicator is currently not well-defined (Tier III). As of the latest report from March 2021, 136 indicators have regular data from at least 50% of countries, 107 indicators have sporadic

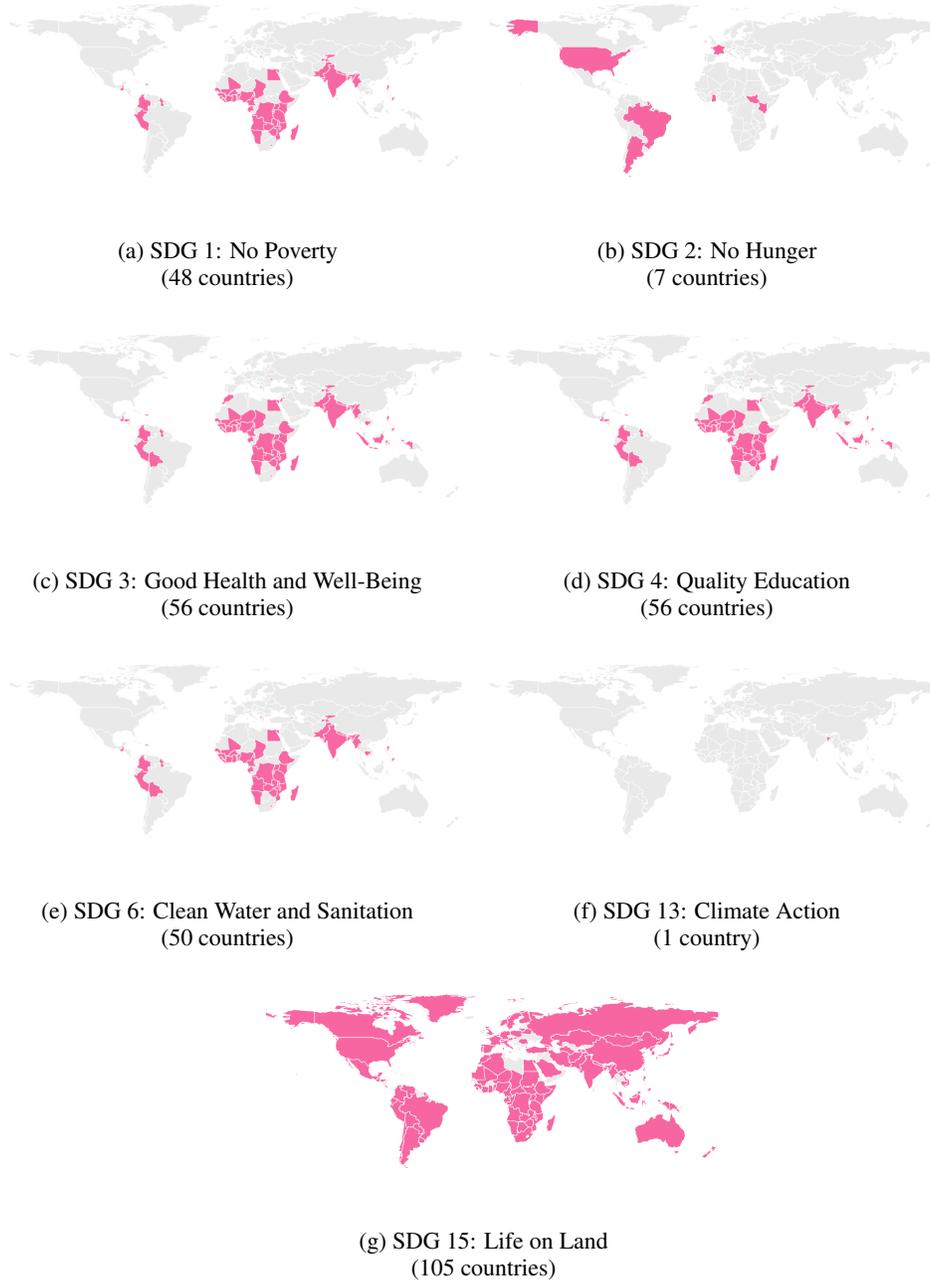


Figure A1: Maps of geographic SUSTAINBENCH coverage per SDG.

data, and 4 indicators are a mix depending on the data of interest (Table A1) [94]. For example, for monitoring global poverty (SDG 1), the proportion of a country's population living below the international poverty line (Indicator 1.1.1) is reported annually for all countries, but the economic loss attributed to natural and man-made disasters (Indicator 1.5.2) is only sparsely documented. We provide descriptions of the 17 Sustainable Development Goals (SDGs) in Table A1.

## D Dataset Details

### D.1 DHS-based datasets

In this section, we detail the process of constructing the poverty, health, education, and water and sanitation labels from DHS surveys. We also give more information about the input imagery that we provide as part of SUSTAINBENCH.

**Labels from DHS survey data** We constructed several indices using survey data from the Demographic and Health Surveys (DHS) program, which is funded by the US Agency for International Development (USAID) and has conducted nationally representative household-level surveys in over 90 countries. For SUSTAINBENCH, we combined survey data covering 56 countries from 179 unique surveys with questions on women’s education, women’s BMI, under 5 mortality, household asset ownership, water quality, and sanitation (toilet) quality. We chose surveys between 1996 (the first year that nightlights imagery is available) and 2019 (the latest year with available DHS surveys)<sup>9</sup> for which geographic data was available. The full list of surveys is shown in Table A3.

- **Asset Wealth Index**

While the SDG indicators define poverty lines expressed in average expenditure (a.k.a. consumption) per day, survey data is much more widely available for household asset wealth than expenditure. Furthermore, asset wealth is considered a less noisy measure of households’ long-run economic well-being [85, 32] and is actively used for targeting social programs [32, 7]. To summarize household-level survey data into a scalar asset wealth index, standard approaches perform principal components analysis (PCA) of survey responses and project them onto the first principal component [31, 85]. The household-level asset wealth index is commonly averaged to create a cluster-level index, where a “cluster” roughly corresponds to a village or local community.

The asset wealth index is built using household asset ownership and infrastructure information as done in prior works [109]. We include the number of rooms used for sleeping in a home (capped at 25); binary indicators for whether the household has electricity and owns a radio, TV, refrigerator, motorcycle, car, or phone (or cellphone); and the quality of floors, water source, and toilet. As “floor type”, “water source type”, and “toilet type” are reported from DHS as descriptive categorical variables (*e.g.*, “piped water”/“flush to pit latrine”), we convert the descriptions to a numeric scale, a standard technique for processing survey data [65]. We use a 1-5 scale where lower numbers indicate the water source is less developed (*e.g.*, straight from a lake) while higher numbers indicate higher levels of technology/development (*e.g.*, piped water); we use a similar 1-5 scale for toilet type and floor type. To calculate the index, we use the first principal component of all the variables mentioned above at a household level, and report the mean at a cluster level. The asset wealth index calculation includes 2,081,808 households total from 87,119 clusters in 48 countries, with a median of 22 households per cluster. Many surveys are dropped because they do not include one of the 12 variables we use to construct the index. The final number of clusters with asset wealth labels in SUSTAINBENCH is only 86,936, as several clusters did not have corresponding satellite imagery inputs. Note that households from these clusters with missing imagery still contributed to the PCA computation, since these clusters were excluded from SUSTAINBENCH only *after* the PCA-based index had already been constructed.

- **Education**

The women’s education metric is created by taking the cluster level mean of “education in single years”. Following [40], we capped the years of education at 18, a common threshold in many surveys which helps avoid outliers. The women’s education metric includes data from 2,910,286 women in 56 countries, with a median of 24 women per cluster.

- **Health**

To create the women’s BMI metric, we first exclude all pregnant women, as the BMI is not adjusted for them. Using the sample of women BMI is appropriate for, we take the cluster

---

<sup>9</sup>Even though a DHS survey may have been conducted over several years, we refer to the “year” of a DHS survey as the year reported for that survey in the DHS Data API: <https://api.dhsprogram.com/>

Table A2: Splits for DHS survey-based tasks. See Table A3 for the mapping between DHS country code and the full country name.

	Train	Validation	Test
DHS Country Codes	30 countries: AL, BD, CD, CM, GH, GU, HN, IA, ID, JO, KE, KM, LB, LS, MA, MB, MD, MM, MW, MZ, NG, NI, PE, PH, SN, TG, TJ, UG, ZM, ZW	13 countries: BF, BJ, BO, CO, DR, GA, GN, GY, HT, NM, SL, TD, TZ	13 countries: AM, AO, BU, CI, EG, ET, KH, KY, ML, NP, PK, RW, SZ
asset wealth index	59,617 examples (69%)	16,776 examples (19%)	10,543 examples (12%)
child mortality rate	69,052 (65%)	17,062 (16%)	19,468 (18%)
women BMI	61,950 (65%)	15,675 (17%)	17,241 (18%)
women education	75,818 (65%)	20,589 (18%)	20,655 (18%)
water index	59,620 (68%)	17,773 (20%)	10,545 (12%)
sanitation index	60,184 (67%)	16,776 (19%)	12,311 (14%)

level mean of reported BMI/100 (as “decimal points are not included” in the DHS data). The women’s BMI metric includes data from 1,781,403 women in 53 countries, with a median of 18 women per cluster.

To create the child mortality metric, we used woman level birth records. For each woman, the DHS reports up to 20 births as well as pregnancy, postnatal care, and health outcomes for each birth. Treating each child (rather than woman) as a record, we keep only the children who were age 5 or younger at the time of survey or who had died (age 5 or younger) no earlier than the year prior to the survey. After identifying the qualifying children, we calculate the number of deaths per 1,000 children by cluster. The child mortality metric includes 1,936,904 children in 56 countries, with a median of 15 children per cluster.

- **Water and Sanitation Indices**

The water and sanitation indices are calculated as the cluster-level mean of our ranking of water quality and toilet type, respectively. The water index calculation includes 2,105,026 households over 49 countries, with a median of 22 households per cluster. The sanitation index calculation includes 2,143,329 households over 49 countries, with a median of 22 households per cluster.

For all indices, we excluded the calculated index for a cluster if fewer than 5 observations are used to create it. For the asset wealth, sanitation, and water indices an observation unit is a household; for the women’s education, BMI and under 5 mortality measures the observation unit is an individual. We also excluded several hundred clusters for which satellite imagery could not be obtained.

For all of the tasks based on DHS survey data, we use a uniform train/validation/test dataset split by country. Delineating by country ensures that there is no overlap between any of the splits—*i.e.*, a model trained on our train split will not have “seen” any part of any image from the test split. The splits are listed in Table A2.

**Multispectral (MS) bands** The main source of inputs for these tasks is satellite imagery, collected and processed in a similar manner as [109]. For each DHS surveyed country and year, we created 3-year median composites of daytime surface reflectance images captured by the Landsat 5, 7, and 8 satellites. Each composite takes the median of each cloud-free pixel available during a 3-year period centered on the year of the DHS survey. (Note the difference from [109], which only chose three distinct 3-year periods for compositing.) As described in [109], the motivation for using 3-year composites is two-fold. First, multi-year median compositing has seen success in similar applications for gathering clear satellite imagery [10], and even in 1-year composites we observed substantial influence of clouds in some regions, given imperfections in the cloud mask. Second, the outcomes that we predict (wealth, health, education, and infrastructure) tend to evolve slowly over time, and we did not want our inputs to be distorted by seasonal or short-run variation. These daytime images

Table A3: 179 DHS surveys from 56 countries spanning 1996-2019 were used to create labels.

DHS Code - Country	Survey IDs (SurveyId field from the DHS Data API)
AL - Albania	AL2008DHS, AL2017DHS
AM - Armenia	AM2010DHS, AM2016DHS
AO - Angola	A02006MIS, A02011MIS, A02015DHS
BD - Bangladesh	BD2000DHS, BD2004DHS, BD2007DHS, BD2011DHS, BD2014DHS, BD2017DHS
BF - Burkina Faso	BF1999DHS, BF2003DHS, BF2010DHS, BF2014MIS, BF2017MIS
BJ - Benin	BJ1996DHS, BJ2001DHS, BJ2012DHS, BJ2017DHS
BO - Bolivia	B02008DHS
BU - Burundi	BU2010DHS, BU2012MIS, BU2016DHS
CD - Congo Democratic Republic	CD2007DHS, CD2013DHS
CI - Cote d'Ivoire	CI1998DHS, CI2012DHS
CM - Cameroon	CM2004DHS, CM2011DHS, CM2018DHS
CO - Colombia	C02010DHS
DR - Dominican Republic	DR2007DHS, DR2013DHS
EG - Egypt	EG2000DHS, EG2003DHS, EG2005DHS, EG2008DHS, EG2014DHS
ET - Ethiopia	ET2000DHS, ET2005DHS, ET2011DHS, ET2016DHS, ET2019DHS
GA - Gabon	GA2012DHS
GH - Ghana	GH1998DHS, GH2003DHS, GH2008DHS, GH2014DHS, GH2016MIS, GH2019MIS
GN - Guinea	GN1999DHS, GN2005DHS, GN2012DHS, GN2018DHS
GU - Guatemala	GU2015DHS
GY - Guyana	GY2009DHS
HN - Honduras	HN2011DHS
HT - Haiti	HT2000DHS, HT2006DHS, HT2012DHS, HT2016DHS
IA - India	IA2015DHS
ID - Indonesia	ID2003DHS
JO - Jordan	J02002DHS, J02007DHS, J02012DHS, J02017DHS
KE - Kenya	KE2008DHS, KE2014DHS, KE2015MIS
KH - Cambodia	KH2000DHS, KH2005DHS, KH2010DHS, KH2014DHS
KM - Comoros	KM2012DHS
KY - Kyrgyz Republic	KY2012DHS
LB - Liberia	LB2007DHS, LB2009MIS, LB2011MIS, LB2013DHS, LB2016MIS, LB2019DHS
LS - Lesotho	LS2004DHS, LS2009DHS, LS2014DHS
MA - Morocco	MA2003DHS
MB - Moldova	MB2005DHS
MD - Madagascar	MD1997DHS, MD2008DHS, MD2011MIS, MD2013MIS, MD2016MIS
ML - Mali	ML1996DHS, ML2001DHS, ML2006DHS, ML2012DHS, ML2015MIS, ML2018DHS
MM - Myanmar	MM2016DHS
MW - Malawi	MW2000DHS, MW2004DHS, MW2010DHS, MW2012MIS, MW2014MIS, MW2015DHS, MW2017MIS
MZ - Mozambique	MZ2009AIS, MZ2011DHS, MZ2015AIS, MZ2018MIS
NG - Nigeria	NG2003DHS, NG2008DHS, NG2010MIS, NG2013DHS, NG2015MIS, NG2018DHS
NI - Niger	NI1998DHS
NM - Namibia	NM2000DHS, NM2006DHS, NM2013DHS
NP - Nepal	NP2001DHS, NP2006DHS, NP2011DHS, NP2016DHS
PE - Peru	PE2000DHS, PE2004DHS, PE2007DHS, PE2009DHS
PH - Philippines	PH2003DHS, PH2008DHS, PH2017DHS
PK - Pakistan	PK2006DHS, PK2017DHS
RW - Rwanda	RW2005DHS, RW2008DHS, RW2010DHS, RW2015DHS
SL - Sierra Leone	SL2008DHS, SL2013DHS, SL2016MIS, SL2019DHS
SN - Senegal	SN1997DHS, SN2005DHS, SN2008MIS, SN2010DHS, SN2012DHS, SN2015DHS, SN2017DHS, SN2018DHS
SZ - Eswatini	SZ2006DHS
TD - Chad	TD2014DHS
TG - Togo	TG1998DHS, TG2013DHS, TG2017MIS
TJ - Tajikistan	TJ2012DHS, TJ2017DHS
TZ - Tanzania	TZ1999DHS, TZ2007AIS, TZ2010DHS, TZ2012AIS, TZ2015DHS, TZ2017MIS
UG - Uganda	UG2000DHS, UG2006DHS, UG2009MIS, UG2011DHS, UG2014MIS, UG2016DHS, UG2018MIS
ZM - Zambia	ZM2007DHS, ZM2013DHS, ZM2018DHS
ZW - Zimbabwe	ZW1999DHS, ZW2005DHS, ZW2010DHS, ZW2015DHS



Figure A2: An example of an input satellite image for the DHS survey-based datasets. This image is of cluster 969 from the 2004 DHS survey of Peru, located at latitude and longitude coordinates of (-12.597851, -69.185416). The left image shows the RGB channels from Landsat surface reflectance. The right image shows the Nightlights band from DMSP.

have a spatial resolution of 30 m/pixel with seven bands which we refer to as the multispectral (MS) bands: RED, GREEN, BLUE, NIR (Near Infrared), SWIR1 (Shortwave Infrared 1), SWIR2 (Shortwave Infrared 2), and TEMP1 (Thermal).

**Nightlights (NL)** We also include nighttime lights (“nightlights”) imagery, using the same sources as [109]. No single satellite captured calibrated nightlights for all of 1996-2019, so we collected DMSP-OLS Radiance Calibrated Nighttime Lights [46] for the years 1996-2011, and VIIRS Nighttime Day/Night Band [29] for the years 2012-2019. DMSP nightlights have 30 arc-second/pixel resolution and are considered unitless, whereas VIIRS nightlights have 15 arc-second/pixel resolution and units of radiance ( $\text{nW cm}^{-2} \text{sr}^{-1}$ ). For the DMSP calibrated nightlights, which only exists as annual composites for a few specific years, we chose the annual composite closest to the year of the DHS survey; furthermore, we use the inter-satellite calibration procedure from [46] to ensure that the DMSP values are comparable across time (a procedure which [109] did not follow). For VIIRS, which provides monthly composites, we perform 3-year median compositing similar to the Landsat images, taking the median of each monthly average radiance over a 3-year period centered on the year of the DHS survey. All nightlights images are resized using nearest-neighbor upsampling to cover the same spatial area as each Landsat image.

The MS and NL satellite imagery were processed in and exported from Google Earth Engine [39]. For each cluster from a given DHS surveyed country-year, we provide one  $255 \times 255 \times 8$  image (7 MS bands, 1 NL band) centered on the cluster’s geocoordinates at a scale of 30 m/pixel. See Figure A2 for an example of an image in our dataset. In our released code, we provide the mean and standard deviation of each band across the entire dataset for input normalization.

The exact image collections we used on Google Earth Engine are as follows:

- USGS Landsat 5, Collection 1 Surface Reflectance Tier 1: LANDSAT/LT05/C01/T1\_SR
- USGS Landsat 7, Collection 1 Surface Reflectance Tier 1: LANDSAT/LE07/C01/T1\_SR
- USGS Landsat 8, Collection 1 Surface Reflectance Tier 1: LANDSAT/LC08/C01/T1\_SR
- DMSP-OLS Global Radiance-Calibrated Nighttime Lights Version 4:  
NOAA/DMSP-OLS/CALIBRATED\_LIGHTS\_V4
- VIIRS Nighttime Day/Night Band Composites Version 1:  
NOAA/VIIRS/DNB/MONTHLY\_V1/VCMSFG

For future releases of SUSTAINBENCH, we would like to update all of the Landsat imagery to the newer “Collection 2” products. New Collection 1 products will not be released beyond January 1, 2022, so we would not be able to use the existing Collection 1 imagery source for future DHS surveys. We would also like to update the VIIRS imagery to the official annual composites released by the Earth Observation Group. We did not provide such imagery in SUSTAINBENCH because they were not available on Google Earth Engine at the time SUSTAINBENCH was compiled.

**Mapillary Images** Mapillary [71] provides a platform for crowd-sourced, geo-tagged street-level imagery. It provides an API to access data such as images, map features, and object detections, automatically blurring faces of human subjects and license plates [72] and allowing users who upload



Figure A3: An example of an input street-level image from Mapillary for the DHS survey-based datasets. The left image is from cluster 10 of Armenia located at (40.192860, 44.515051). The right image is from cluster 92 of Benin, located at (2.347327, 6.402679).

images to manually blur if any are missed [3] for privacy. We retrieved only images that intersect with a DHS cluster. A given image must satisfy two conditions to intersect with a DHS cluster: 1) its geo-coordinates must be within 0.1 degree latitude and longitude to the cluster’s geo-location, and 2) it must have been captured within 3 years before or after the year of the DHS datapoint. Each image has metadata, including a unique ID, timestamp of capture in milliseconds, year of capture, latitude, and longitude. All Mapillary images have 3 channels (RGB), and the length of the shorter side is 1024.

22,052 DHS clusters (18.7% of all clusters included in SUSTAINBENCH), spanning 48 countries, have a non-zero number of Mapillary images. Of these clusters with Mapillary images, the number of images per cluster ranges from 1 to a maximum of 300, with a mean of 76.3 and median of 94.0. A total of 1,682,613 Mapillary images are included in SUSTAINBENCH. Figure A3 shows some example Mapillary images.

**Comparison with Related Works** Table A5 summarizes the related works for the DHS-based tasks in SUSTAINBENCH.

As shown in Table A4, the DHS-based datasets in SUSTAINBENCH build on the previous works of Jean et al. 52 and Yeh et al. 109, which pioneered the application of computer vision on satellite imagery to estimate a cluster-level asset wealth index. Notably, for the task of predicting poverty over space, SUSTAINBENCH’s dataset is nearly  $5\times$  larger than the dataset included in [109] (over  $2\times$  the number of countries, and  $3\times$  the temporal coverage). Our dataset also has advantages over other related works which often rely on proprietary imagery inputs [52, 45, 36], are limited to a small number of countries [12, 30, 64, 36, 105], or have coarser label resolution [73]. Other researchers have explored using non-imagery inputs for poverty prediction, including Wikipedia text data [87] and cell phone records [15]; while such multi-modal data are not currently in SUSTAINBENCH, we are considering including them in future versions.

For the non-poverty tasks pertaining to health, education, and water/sanitation, there are extremely few ML-friendly datasets. Head et al. 45 comes closest to SUSTAINBENCH in having predicted similar indicators (women BMI, women education, and clean water) derived from DHS survey data. Also, like us, their results suggest that satellite imagery may be less accurate at predicting these non-poverty labels in developing countries. However, because they used proprietary imagery inputs, their dataset is not accessible and cannot serve as a public benchmark. A large collaborative effort [65] gathered survey and census data for creating clean water and sanitation labels in over 80 countries, but they did not provide satellite imagery inputs and only publicly released outputs of their geostatistical model, not the labels themselves. Again, SUSTAINBENCH has significant advantages over other related works that use proprietary data [45, 36, 67], are limited to a small number of countries [36, 64], or do not publicly release their labels [65].

**Dataset Impact** Most low-income regions lack data on income and wealth at fine spatial scales. Even at coarse spatial scales, temporal resolution can still be bad; Figure 1 in Burke et al. 20 shows that, in some countries, as many as two decades can pass between successive nationally representative economic surveys. Inferring economic welfare from satellite or street-level imagery offers one solution to the lack of surveys.

Table A4: Comparison of related datasets using satellite images to predict DHS asset wealth index. \*The clusters in SUSTAINBENCH are a superset of the clusters included in [109] except for 2 clusters that had fewer than the minimum of 5 observations we required for inclusion in SUSTAINBENCH.

	<b>Jean et al. (2016) [52]</b>	<b>Yeh et al. (2020) [109]</b>	<b>SUSTAINBENCH</b>
Geographic range	5 countries in Africa	23 countries in Africa	56 countries in 6 continents
Temporal range	2010-2013	2009-2016	1996-2019
Dataset size	3,034 clusters	19,669 clusters	86,936 clusters*
Labels	asset wealth index with different asset variables in PCA for each country	asset wealth index with PCA pooled over 30 countries (a superset of the 23 countries with provided imagery)	asset wealth index with PCA pooled over all 56 countries
Daytime satellite imagery	~2.5m/px Google Static Maps daytime images, 3 bands, proprietary license	30m/px resolution, 7 bands, Landsat 5/7/8 surface reflectance 3-year median composites (binned to either 2009-11, 2012-14, or 2015-17), some cloud masking	30m/px resolution, 7 bands, Landsat 5/7/8 surface reflectance 3-year median composites (centered on survey year), improved cloud masking
Nightlights	~1km/px DMSP-OLS Nighttime Lights (uncalibrated), annual composite chosen to match survey year	(2009-2011) ~1km/px DMSP-OLS Radiance-Calibrated Nighttime Lights, without inter-satellite calibration, 3-year composite; (2012-2017) ~500m/px VIIRS Stray Light Corrected Nighttime Day/Night Band, 3-year median composite of monthly images	(1996-2011) ~1km/px DMSP-OLS Radiance-Calibrated Nighttime Lights, with inter-satellite calibration, annual composite chosen closest to survey year; (2012-2019) ~500m/px VIIRS Nighttime Day/Night Band (these are higher quality than the stray light corrected images), 3-year median composite of monthly images

Indeed, many governments turned to ML-based poverty mapping techniques during the COVID-19 pandemic to identify and prioritize vulnerable populations for targeted aid programs. For example, the government of Togo wanted to send aid to over 500,000 vulnerable people impacted by the pandemic. But like most low-income countries, Togo lacks accurate data on income and wealth at fine spatial scales. Working with a research group at UC Berkeley [6, 14], the government was able to quickly deploy ML-based poverty mapping methods with satellite imagery inputs in order to identify who needs aid the most and then target cash payments to them. Likewise, the governments of Nigeria [66], Mozambique, Liberia, and the Democratic Republic of the Congo [38] also used satellite imagery analysis for identifying and prioritizing neighborhoods with vulnerable individuals for their targeted social protection programs.

Finally, we highlight how ML-based poverty maps can feed into other policy evaluations. Researchers recently combined longitudinal ML-generated poverty maps of rural Uganda with data on expansion of the electric grid. By applying causal inference approaches, they were able to infer the impact of electrification on local livelihoods [78]. This work presents a scalable technique for measuring the effectiveness of large-scale infrastructure investments.

Table A5: Non-exhaustive comparison of related works and datasets for predicting DHS-based labels from satellite imagery, street-level imagery, or other non-survey inputs. “None” indicates that, to the best of our knowledge, we are not aware of existing works that predict the DHS label at scale. “SB” is short for SUSTAINBENCH. (While [65] uses survey data as inputs, they generate a prediction map including for locations where survey data were not available.)

	Satellite imagery	Street-level imagery	Other inputs
<b>poverty</b> SB includes 56 countries	[52] (5 countries) [109] (23 countries) [73] (37 countries) [45] (4 countries) [12] (Mexico) [30] (Sri Lanka) [105] (Kenya)	[64] (2 countries) [36] (USA)	[87] (Wikipedia text, 31 countries) [15] (phone records, Rwanda)
<b>women BMI</b> SB includes 53 countries	[45] (4 countries) [67] (USA)	[64] (India)	none
<b>child mortality</b> SB includes 56 countries	none	none	none
<b>women education</b> SB includes 56 countries	[45] (4 countries) [112] (9 countries)	[36] (USA)	none
<b>clean water</b> SB includes 49 countries	[45] (4 countries)	none	[65] (survey data, 88 countries)
<b>sanitation</b> SB includes 49 countries	none	none	[65] (survey data, 89 countries)



Figure A4: An example of a pair of satellite imagery inputs for predicting change in poverty over time for the Nigeria cluster located at (7.797380, 4.778803), in (a) 2010 and (b) 2015. Landsat RGB bands (left) and the DMSP/VIIRS nightlights band (right) are shown for each year.

## D.2 Data for Predicting Change in Poverty Over Time

The task of predicting change in poverty over time uses labels calculated from household surveys conducted by the World Bank’s Living Standards Measurement Study (LSMS) program. The LSMS surveys are similar to the DHS surveys described in the previous section. However, unlike DHS surveys, LSMS provides panel data—*i.e.*, the same households are surveyed over time, facilitating comparison over time.

We start by compiling the same survey variables from the DHS asset index, except for refrigerator ownership because it is not included in the LSMS Uganda survey. (See the previous section for details on the survey variables included for the DHS asset index.) As with the DHS asset index, we convert “floor type”, “water source type”, and “toilet type” variables from descriptive categorical variables to a 1-5 ranked scale.

Based on the panel survey data, we calculate two PCA-based measures of change in asset wealth over time for each household: `diffOfIndex` and `indexOfDiff`. For `diffOfIndex`, we first assign each household-year an asset index computed as the first principal component of all the asset variables; this is the same approach used for the DHS asset index. Then, for each household, we calculate the difference in the asset index across years, which yields a “change in asset index” (hence the name `diffOfIndex`). In contrast, `indexOfDiff` is created by first calculating the difference in asset variables in households across pairs of surveys for each country and then computing the first principal component of these differences; for each household, this yields a “index of change in assets” across years (hence the name `indexOfDiff`). These measures are then averaged to the cluster-level to create cluster-level labels. We excluded a cluster if it contained fewer than 3 surveyed households.

As an example, consider an Ethiopian household  $h$  that is surveyed in 2011 and 2015. This household would have 2 labels:

$$\begin{aligned} \text{diffOfIndex}(h, 2011, 2015) &= \text{assetIndex}(h, 2015) - \text{assetIndex}(h, 2011) \\ \text{indexOfDiff}(h, 2011, 2015) &= \text{firstPrincipalComponent}(\text{assets}(h, 2015) - \text{assets}(h, 2011)) \end{aligned}$$

If the set  $\mathcal{C}$  of households represents a cluster in Ethiopia, then its cluster-level labels are

$$\begin{aligned} \text{diffOfIndex}(\mathcal{C}, 2011, 2015) &= \frac{1}{|\mathcal{C}|} \sum_{h \in \mathcal{C}} \text{diffOfIndex}(h, 2011, 2015) \\ \text{indexOfDiff}(\mathcal{C}, 2011, 2015) &= \frac{1}{|\mathcal{C}|} \sum_{h \in \mathcal{C}} \text{indexOfDiff}(h, 2011, 2015) \end{aligned}$$

The LSMS-based labels include data for 2,763 cluster-years (comprising 17,215 household-years) from 11 surveys for 5 African countries. Table A6 gives the full list of LSMS surveys used,<sup>10</sup> and

<sup>10</sup>LSMS survey data citations (all data was downloaded from <https://microdata.worldbank.org>):

Central Statistical Agency of Ethiopia. Ethiopia Rural Socioeconomic Survey (ERSS) 2011-2012. Public Use Dataset. Ref: ETH\_2011\_ERSS\_v02\_M. Downloaded on August 25, 2021.

Central Statistical Agency of Ethiopia. Ethiopia Socioeconomic Survey, Wave 3 (ESS3) 2015-2016. Public Use Dataset. Ref: ETH\_2015\_ESS\_v02\_M. Downloaded on August 26, 2021.

National Statistical Office, Government of Malawi. Integrated Household Panel Survey (IHPS) 2010-2013-2016. Public Use Dataset. Ref: MWI\_2010-2016\_IHPS\_v03\_M. Downloaded on September 3, 2021.

National Bureau of Statistics, Federal Republic of Nigeria. Nigeria General Household Survey (GHS), Panel 2010, Wave 1. Ref: NGA\_2010\_GHSP-W1\_v03\_M. Dataset downloaded on September 4, 2021.

National Bureau of Statistics, Federal Republic of Nigeria. Nigeria General Household Survey (GHS), Panel 2015-2016, Wave 3. Ref: NGA\_2015\_GHSP-W3\_v02\_M. Dataset downloaded on September 4, 2021.

Tanzania National Bureau of Statistics (NBS). Tanzania National Panel Survey 2008-2009 (Round 1). Ref: TZA\_2008\_NPS-R1\_v03\_M. Dataset downloaded on September 4, 2021.

Tanzania National Bureau of Statistics (NBS). Tanzania National Panel Survey Report (NPS) - Wave 2, 2010-2011. Dar es Salaam, Tanzania: NBS. Ref: TZA\_2010\_NPS-R2\_v03\_M. Dataset downloaded on September 5, 2021.

Tanzania National Bureau of Statistics (NBS). Tanzania National Panel Survey Report (NPS) - Wave 3, 2012-2013. Dar es Salaam, Tanzania: NBS. Ref: TZA\_2012\_NPS-R3\_v01\_M. Dataset downloaded on September 4, 2021.

Uganda Bureau of Statistics. Uganda National Panel Survey (UNPS), 2005-2009. Public Use Dataset. Ref: UGA\_2005-2009\_UNPS\_v01\_M. Downloaded on August 25, 2021.

Uganda Bureau of Statistics. Uganda National Panel Survey (UNPS), 2013-2014. Public Use Dataset. Ref: UGA\_2013\_UNPS\_v01\_M. Downloaded on August 25, 2021.

Table A6: LSMS surveys

Country and Year	Survey Title	Survey ID
Ethiopia 2011	Rural Socioeconomic Survey 2011-2012	ETH_2011_ERSS_v02_M
Ethiopia 2015	Socioeconomic Survey 2015-2016, Wave 3	ETH_2015_ESS_v03_M
Malawi 2010 & 2016	Integrated Household Panel Survey 2010-2013-2016	MWI_2010-2016_IHPS_v03_M
Nigeria 2010	General Household Survey, Panel 2010-2011, Wave 1	NGA_2010_GHSP-W1_v03_M
Nigeria 2015	General Household Survey, Panel 2015-2016, Wave 3	NGA_2015_GHSP-W3_v02_M
Tanzania 2008	National Panel Survey 2008-2009, Wave 1	TZA_2008_NPS-R1_v03_M
Tanzania 2012	National Panel Survey 2012-2013, Wave 3	TZA_2012_NPS-R3_v01_M
Uganda 2005 & 2009	National Panel Survey 2005-2009	UGA_2005-2009_UNPS_v01_M
Uganda 2013	National Panel Survey 2013-2014	UGA_2013_UNPS_v01_M

Table A7: Number of clusters and households included from each country for the “predicting change in poverty over time” task, based on LSMS survey data.

Country	# clusters	# households
Ethiopia	235	1128
Malawi	101	1085
Nigeria	462	3093
Tanzania	300	1431
Uganda	189	1247
Total	1287	7984

Table A7 gives the number of clusters and households included for each country. See Figure A4 for an example of the satellite imagery inputs.

The labels and inputs provided in SUSTAINBENCH for this task are similar (but not identical) to the labels and inputs used in [109]. While the underlying LSMS survey data used are the same, there are 3 key differences.

1. In SUSTAINBENCH, for each country, we only used data from households that are present in all surveys of that country. In Uganda, for example, we only keep households that were surveyed repeatedly in all of the 2005, 2009, and 2013 surveys. This is different from [109] which included any household that was present in two survey years—*e.g.*, a household in Uganda 2005 and Uganda 2009, but not Uganda 2013.
2. The recoding of the floor, water, and toilet quality variables was made more consistent across countries and now closely matches the ranking introduced in [65].
3. As in the case of the DHS-based datasets, the satellite imagery inputs have been improved. See Table A4 for details.

**Comparison with Related Works** To the best of our knowledge, the LSMS-based poverty change over time dataset in SUSTAINBENCH and its predecessor in [109] are the only datasets specifically designed as an index of asset wealth change. For related works on mapping poverty, see the “Comparison with Related Works” for DHS-based tasks in Appendix D.1.

### D.3 Cropland Mapping with Landsat

We release a dataset for performing weakly supervised classification of cropland in the United States using the data from Wang et al. 102, which has not been released previously. While densely segmented labels are time-consuming and infeasible to generate for a region as large as Sub-Saharan Africa, pixel-level and image-level labels are often already available and much easier to create. Figure A5 shows an example from the dataset.

The study area spans from 37°N to 41°30’N and from 94°W to 86°W, and covers an area of over 450,000km<sup>2</sup> in the United States Midwest. We chose this region because the US Department of Agriculture (USDA) maintains high-quality pixel-level land cover labels across the US [69], allowing

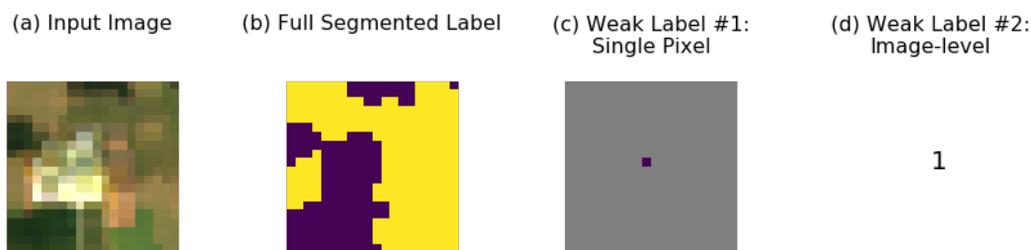


Figure A5: An example from the cropland mapping dataset [102], showing (a) an example Landsat image, (b) its corresponding fully segmented label, (c) single pixel weak label, and (d) image-level weak label.

us to evaluate the performance of algorithms. Land cover-wise, the study region is 44% cropland and 56% non-crop (mostly temperate forest).

The Landsat Program is a series of Earth-observing satellites jointly managed by the USGS and NASA. Landsat 8 provides moderate-resolution (30m) satellite imagery in seven surface reflectance bands (ultra blue, blue, green, red, near infrared, shortwave infrared 1, shortwave infrared 2) designed to serve a wide range of scientific applications. Images are collected on a 16-day cycle.

We computed a single composite by taking the median value at each pixel and band from January 1, 2017 to December 31, 2017. We used the quality assessment band delivered with the Landsat 8 images to mask out clouds and shadows prior to computing the median composite. The resulting seven-band image spans 4.5 degrees latitude and 8.0 degrees longitude and contains just over 500 million pixels. The composite was then divided into 200,000 tiles of  $50 \times 50$  pixels each. This full dataset was not released previously with Wang et al. 102.

The ground truth labels from the Cropland Data Layer [69] are at the same spatial resolution as Landsat, so that for every Landsat pixel there is a corresponding {cropland, not cropland} label. For each image, we generate two types of weak labels: (1) single pixel and (2) image-level, both with the goal of generating dense semantic segmentation predictions. The image-level label is  $\in \{\geq 50\% \text{ cropland}, < 50\% \text{ cropland}\}$ .

**Comparison with Related Works** Cropland has already been mapped globally [18, 35] or for the continent of Africa [106] in multiple state-of-the-art land cover maps. However, existing land cover maps are known to have low accuracy throughout the Global South [56]. One reason behind this low accuracy is that existing maps have been created with SVM or tree-based algorithms that take into account a single pixel at a time [18, 35, 106]. Kerner et al. 56 showed that a multi-headed LSTM (still trained on single pixels) outperformed SVM and random forest classifiers on cropland prediction in Togo. Using a larger spatial context, *e.g.*, in a CNN, could lead to further accuracy gains. However, ground label scarcity remains a bottleneck for applying deep learning models to map cropland. Wang et al. 102 showed that weak labels in the form of single pixel or image-level classes can still supervise a U-Net to segment cropland at accuracies better than SVM or random forest classifiers. We release this dataset, which is the first dataset for weakly supervised cropland mapping, as a benchmark for algorithm development. The dataset is in the U.S. Midwest because cropland labels there are of high accuracy; methods developed on this dataset could be paired with newly generated weak labels in low-income regions to generate novel, high-accuracy cropland maps (see below for an example application).

**Dataset Impact** High accuracy cropland mapping in the Global South can have significant impacts on the planning of government programs and downstream tasks like crop type mapping and yield prediction. For instance, during the COVID-19 pandemic, the government of Togo announced a program to boost national food production by distributing aid to farmers. However, the government lacked high-resolution spatial information about the distribution of farms across Togo, which was crucial for designing this program. Existing global land cover maps, despite including a cropland class, were low in accuracy across Togo. The government collaborated with researchers at the University of Maryland to solve this problem, and in Kerner et al. 56 the authors created a high-resolution map of cropland in Togo for 2019 in under 10 days. The authors pointed out that this

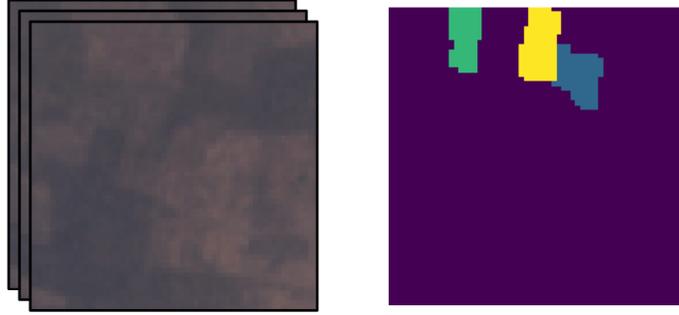


Figure A6: An example from the crop type mapping dataset [83]. The left image represents a satellite image timeseries (figure displays PlanetScope imagery) and the right image represents a segmentation map.

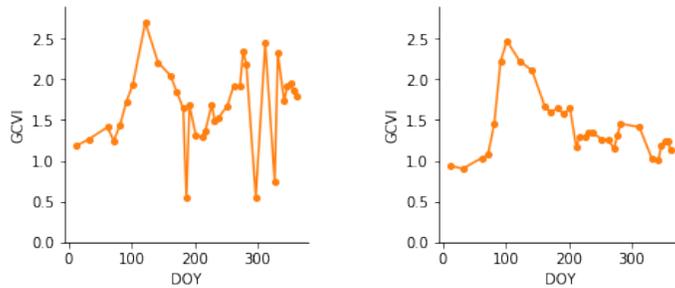


Figure A7: Example time series of the GCVI band computed from Sentinel-2 satellite bands [58], after clouds were masked out. Both examples happen to be of the crop type “Cassava”.

case study demonstrates “a successful transition of machine learning research to operational rapid response for a real humanitarian crisis” [56].

#### D.4 Crop Type Mapping with Planet and Sentinel Imagery

As introduced in [83], these datasets contain satellite imagery from Ghana and South Sudan. Sentinel 1 (10m resolution), Sentinel 2 (10m resolution), and Planet’s PlanetScope (3m resolution) time series imagery are used as inputs for this task. As described in [83], Planet imagery is incorporated to help mitigate issues from high cloud cover and small field sizes. We include three S1 bands (VV, VH, VH/VV), ten S2 bands (blue, green, red, near infrared, four red edge bands, two short wave infrared bands), and all four PlanetScope bands (blue, green, red, near infrared). We also construct normalized difference vegetation index (NDVI) and green chlorophyll vegetation index (GCVI) bands for PlanetScope and S2 imagery.

Ground truth labels consist of a 64x64 pixel segmentation map, with each pixel containing a crop label. Ghana locations are labeled for Maize, Groundnut, Rice, and Soya Bean, while South Sudan locations are labeled for Sorghum, Maize, Rice, and Groundnut.

**Comparison with Related Works** SUSTAINBENCH’s crop type datasets and existing crop type datasets are summarized in Table A8. A version of SUSTAINBENCH’s Ghana/South Sudan dataset was released previously and is currently housed on Radiant MLHub. We highlight key differences between SUSTAINBENCH’s dataset and the one used in Rustowicz et al. 83. We use the same train, validation, and test splits used in [83], though we use the full 64x64 imagery provided, while [83] further subdivided imagery into 32x32 pixel grids due to memory constraints. We also include variable length time series with zero padding and masking, while [83] trimmed the respective time series down to the same length. We include variable length time series with the reasoning that future research should be extendable to variable length time-series imagery. The metrics cited in Table 2 are on the original Rustowicz et al. 83 dataset.

Table A8: A comparison of SUSTAINBENCH’s crop type datasets with existing datasets. A dataset is only included if it is designed for crop type mapping, is publicly available, and provides both inputs and outputs in ML-friendly formats. Compared to Table 1, we include datasets that lack train/test splits and standardized benchmarks, though we make a note of their existence in the columns.

Dataset Collection	Dataset #	Geography	Time	Inputs	Size	Small-holder?	Data splits?	Base-line?
SUSTAINBENCH	1 [83]	Ghana and South Sudan	2016-17	Sat. image time series	4,439 and 837 fields	✓	✓	✓
	2 [54, 58]	Kenya	2017	Sat. time series	5,746 fields	✓	✓	✓
Radiant MLHub [77]	1 [83]	Ghana and South Sudan	2016-17	Sat. image time series	4,439 and 837 fields	✓	✓	✓
	2 [55]	Kenya	2019	Sat. image time series	4,668 fields	✓	✓	✓
	3	Kenya	2019	Sat. image time series	319 fields	✓		
	4	Tanzania	2019	Sat. image time series	392 fields	✓		
	5	Uganda	2017	Sat. image time series	232 fields	✓		
	6 [22]	Rwanda	2018-19	Drone imagery	2,611 points	✓	✓	✓
	7 [79]	Uzbekistan and Tajikistan	2015-18	Sat. imagery time series	8,435 fields			
	8	South Africa	2017-18	Sat. imagery time series	Unknown		✓	✓

## D.5 Crop Type Mapping with Sentinel-2 Time Series

The data from Jin et al. 54 and Kluger et al. 58 comes from three regions in Kenya: Bungoma, Busia, and Siaya. They use time series from the multi-spectral Sentinel-2 (10m resolution) to differentiate crop types at individual pixels in the fields (Figure A7). Time series span from January 1, 2017 to December 31, 2017. All 13 Sentinel-2 bands were used as features, along with GCVI (green chlorophyll vegetation index) as a fourteenth band. Cloudy observations were removed using the QA60 band delivered with Sentinel-2 and the Hollstein Quality Assessment measure.

Ground truth labels are from a survey conducted on crop types during the long rains season in Kenya in 2017. The labels span 9 crop types: Sweet Potatoes, Cassava, Maize, Banana, Beans, Groundnut, Sugar Cane, Other, and Non-crop.

The train, validation, and test sets are split by region to encourage discovery of features and development of methods that generalize across regions. One region is the training and validation region, while the other two regions are test regions.

**Comparison with Related Works** SUSTAINBENCH’s crop type datasets and existing crop type datasets are compared in Table A8. A dataset was only included in the table if it is publicly available and provides inputs and outputs in ML-friendly formats. There is considerable work underway in the remote sensing community, led by the Radiant Earth Foundation, to collect and disperse crop type data to improve the state-of-the-art classification. SUSTAINBENCH’s crop type dataset in Kenya complements existing datasets. It is one of the largest available crop type datasets in a smallholder system. It also has defined train/val/test splits and baselines, which not all public crop type datasets do. One of the train/val/test split options is also designed to test model generalizability across geography by splitting along geographic clusters, which no other datasets do. We recommend that ML researchers test their methods on as many available datasets as possible to ensure model generalizability.

**Dataset Impact** The crop type labels that we released in Kenya were the same labels used to create the first-ever maize classification and yield map across that entire country [54]. Kenya is one of the largest maize producers in sub-Saharan Africa, and studying maize production there could improve food security in the region. Jin et al. 54 used a random forest trained on seasonal median composites of satellite imagery to predict maize with an accuracy of only 63%. It is worth investigating how other machine learning models using a year’s full time series could improve on this. As an example

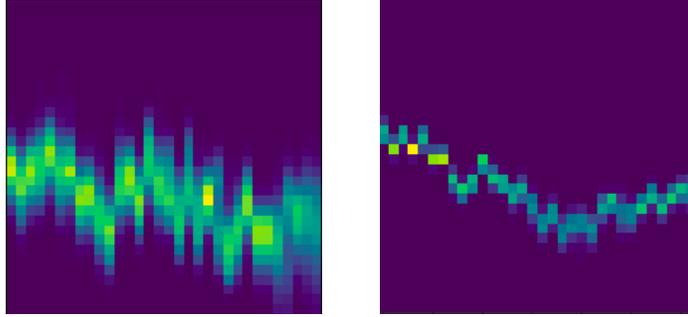


Figure A8: An example from the MODIS crop yield dataset. The spectral histograms are over the 2015 harvest season in the La Capital department, Santa Fe province, Argentina, with a soybean yield of 2.947 metric tonnes per hectare. The left image shows surface reflectance band 5 out of 7, covering wavelengths from 1230-1250nm. The right image shows surface temperature band 1, covering daytime land surface temperatures.

of novel insights resulting from one of our datasets: analysis of the maize yield map in Jin et al. 54 revealed that 72% of variation in predicted maize yields could be explained by soil factors, suggesting that increasing nitrogen fertilizer application should be a priority for increasing smallholder yields in Kenya.

## D.6 Crop Yields and MODIS

These datasets are constructed as an expansion of the dataset used in [101]. They are created using Moderate Resolution Imaging Spectroradiometer (MODIS) satellite imagery, which is freely accessible via Google Earth Engine and provides coverage of the entire globe. Specifically, we use 8-day composites of MODIS images to get 7 bands of surface reflectance at different wavelengths (3 visible and 4 infrared bands) from the MOD09A1 [97] collection, 2 bands of day and night surface temperatures from MYD11A2 [100], and a land cover mask from MCD12Q1 [34] to distinguish cropland from other land. For each of the 9 bands of reflectance and temperature imagery and each of the 32 timesteps within a year’s harvest season, we bin pixel values into 32 ranges, giving a  $32 \times 32 \times 9$  final histogram. We create one such dataset for each of Argentina, Brazil, and the United States, with 9049 datapoints for the United States, 1615 for Argentina, and 384 for Brazil.

The ground truth labels are the regional crop yield per harvest, in metric tonnes per cultivated hectare, as collected from Argentine Undersecretary of Agriculture [8], the Brazilian Institute of Geography and Statistics [17], and the United States Department of Agriculture [95].

**Comparison with Related Works** SUSTAINBENCH releases the crop yield datasets from two previous works [110, 101] for the first time. To date, very few crop yield datasets exist, because yields require expensive farm survey techniques (*e.g.*, crop cuts) to measure. The datasets that do contain field-level yields are privately held by researchers, government agencies, or NGOs. SUSTAINBENCH’s datasets therefore provide yields at the county level. Furthermore, crop yield prediction is challenging as it requires processing a temporal sequence of satellite images. We provide ML-friendly inputs in the form of histograms of weather and satellite features over each county.

**Dataset Impact** Tracking crop yields is crucial to measuring agricultural development and deciding resource allocation, with downstream applications to food security, price stability, and agricultural worker income. Notably, most developed countries invest in forecasting and tracking crop yield. For example, the European Commission JRC’s crop yield forecasts and crop production estimates inform the EU’s Common Agricultural Policy and other agricultural programs [1]. By involving satellite images in the crop yield prediction process, we aim to make timely predictions available in developing countries where ground surveys are costly and infrequent. Furthermore, we provide satellite histograms rather than human-engineered indices like NDVI, which are more human-friendly for visualization but discard a significant amount of potentially-relevant information. In doing so, we

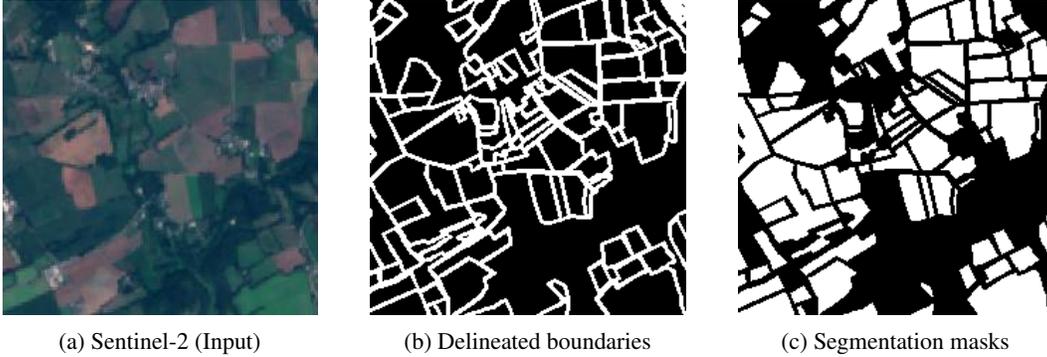


Figure A9: An example from the field delineation dataset [9]. From left to right, an input Sentinel-2 image, its corresponding delineated boundaries, and its corresponding segmentation masks.

hope to encourage the development of ML techniques that make use of more complete and useful features to generate better predictions.

#### D.7 Field Delineation with Sentinel-2

As introduced in [9], the dataset consists of Sentinel-2 satellite imagery in France<sup>11</sup> over the 3 time ranges January-March, April-June, and July-September in 2017. The image has resolution  $224 \times 224$  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 1572 training samples, 198 validation samples, and 196 test samples. We use a different data split from Aung et al. 9 to remove overlapping between the train, validation and test split. An example of the dataset is shown in Figure A9.

**Comparison with Related Works** To our knowledge, SustainBench has released the first public field boundary dataset with satellite image inputs and ML-friendly outputs. That is, some countries in Europe (*e.g.*, France) have made vector files of field boundaries public on their government websites, but without corresponding satellite imagery inputs or raster field boundary outputs. We provide these inputs and outputs. While field segmentation datasets from the U.S., South Africa, and Australia were used in prior field delineation research [107, 98, 99], none of those datasets are publicly available. We are also currently working on collecting field boundaries in low-income countries, but this data will be added to SustainBench at a later date, not in time for this submission.

**Dataset Impact** Automated field delineation makes it easier for farmers to access field-level analytics; previously, manual boundary input was a major deterrent from adopting digital agriculture [98]. Digital agriculture can improve yields while minimizing the use of inputs like fertilizer that cause environmental pollution – with the net effect of increasing farmer profit. The development of a new attention-based neural network architecture (called FracTAL ResUNet) enabled the delineation of 1.7 million fields in Australia from satellite imagery [99]. These field boundaries have since been produced by CSIRO, the Australian government agency for scientific research. This is an example where a novel deep learning architecture enabled the creation of operational products in agriculture. However, the Australia dataset is not publicly available. Our goal is for the release of SUSTAINBENCH’s field boundary dataset in France to enable further architecture development and identify which model works best for field delineation.

#### D.8 Brick Kiln Detection with Sentinel-2

Brick manufacturing is a major source of pollution in South Asia, but the industry is largely comprised of small-scale, informal producers, making it difficult to monitor and regulate. Identifying brick kilns automatically from satellite imagery can help improve compliance with environmental regulations

<sup>11</sup><https://www.data.gouv.fr/en/datasets/registre-parcellaire-graphique-rpg-contours-des-parcelles-et-ilots-cultureaux-et-leur-groupe-de-cultures-majoritaire/>



Figure A10: An example of Sentinel-2 satellite imagery for brick kiln classification. On the left is a positive example of an image showing a brick kiln, while the right image is a negative example (*i.e.*, no brick kiln).

and measure their impact on the health of nearby populations. We provide Sentinel-2 satellite imagery at 10m/pixel resolution available through Google Earth Engine [39]. The images have size  $64 \times 64 \times 13$ px, where the order of the bands correspond to the bands B1 through B12 on the Earth Engine Data Catalog, where B2 is Blue, B3 is Green, and B4 is Red. The other bands include aerosols, color infrared, short-wave infrared, and water vapor data.

**Comparison with Related Works** A recent study detected brick kilns from high-resolution (1m/pixel) satellite imagery and hand-validated the predictions, providing ground truth locations of brick kilns in Bangladesh for the time period of October 2018 to May 2019 [63]. The imagery could not be shared publicly because they were proprietary. Hence, we provide Sentinel-2 satellite imagery instead. With help from domain experts, we verified the labels of each image as not containing a brick kiln (class 0) or containing a brick kiln (class 1) based on the ground truth locations provided by [63]. There were roughly 374,000 examples total, with 6329 positives. We sampled 25% of the remaining negatives, removed any null values, and included the remaining 67,284 negative examples in our dataset.

**Dataset Impact** SUSTAINBENCH introduces the first publicly released dataset of this size and quality on detecting brick kilns across Bangladesh from satellite imagery. This dataset was manually labeled and verified in-house by domain experts. Brick kiln detection is a challenging task because of the sparsity of kilns and lack of similar training data, but with recent developments in satellite monitoring [63], it plays a key role in affecting policy developed by public health experts, industry stakeholders (*e.g.*, kiln owners), and government agencies [88]. SUSTAINBENCH is the first to contribute a large dataset for this task, and the results of models will be utilized by policymakers.

## D.9 Representation Learning for Land Cover Classification

The dataset from Jean et al. 53 uses imagery from the USDA’s National Agriculture Imagery Program (NAIP), which provides aerial imagery for public use that has four spectral bands (red (R), green (G), blue (B), and infrared (N)) at 0.6 m ground resolution. They obtained an image of Central Valley, California near the city of Fresno for the year 2016, spanning latitudes [36.45, 37.05] and longitudes [-120.25, -119.65]. There are over 12 billion pixels in the dataset.

The Cropland Data Layer (CDL) is a raster georeferenced land cover map collected by the USDA for the continental United States [69] and serves as ground truth labels of land cover. Offered at 30 m resolution, CDL includes 132 class labels spanning crops, developed areas, forest, water, and more. In the NAIP dataset over Central Valley, CA, 66 CDL classes are observed. CDL is used as ground truth for evaluation by upsampling it to NAIP resolution and taking the mode over each NAIP image.

**Comparison with Related Works** Representation learning on natural images often uses canonical computer vision datasets like ImageNet and Pascal VOC to evaluate new methods. Satellite imagery lacks an analogous dataset. The high-resolution aerial imagery dataset released in SUSTAINBENCH aims to fill this void for land cover mapping with high-resolution inputs in particular. We note that, for object detection or lower resolution inputs, repurposing a dataset like fMoW [23], SpaceNet [96],



Figure A11: Example images from the NAIP dataset collected by Jean et al. 53. The left image is an example of the “Grapes” class and the right image is an example of the “Urban” class.

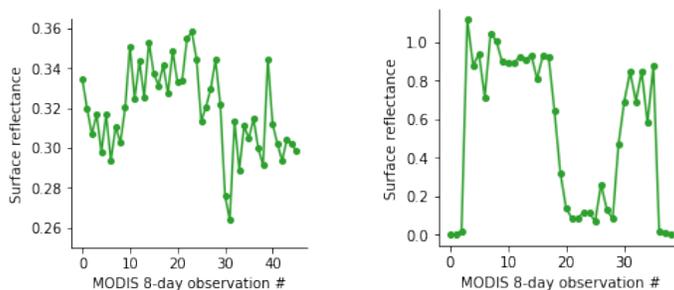


Figure A12: Example time series from the 8-day MODIS satellite product collected by Wang et al. 104. The left time series is an example from Mauritania and the right time series is an example from Canada.

Sen12MS [86], or BigEarthNet [89] would also be appropriate. To our knowledge, such repurposing has not yet been done.

**Dataset Impact** Many tasks in sustainability monitoring have abundant unlabeled imagery but scarce labels. Land cover mapping in low-income regions is one example; crop type mapping in smallholder systems is another. By learning representations of satellite images in an unsupervised or self-supervised way, we may be able to improve performance on SDG-related tasks for the same number of training labels.

#### D.10 Out-of-Domain Land Cover Classification

Wang et al. 104 sampled one thousand  $10\text{km} \times 10\text{km}$  regions uniformly at random from the Earth’s land surface, and removed regions that have fewer than 2 unique land cover classes and regions where one land cover type comprises more than 80% of the region’s area. This resulted in 692 regions across 105 countries. The authors placed the 103 regions from Sub-Saharan Africa into the meta-test set and split the remainder into 485 meta-train and 104 meta-val regions at random. We provide the user with the option of placing any continent into the meta-test set and splitting the other continents’ regions at random between the meta-train and meta-val sets.

In each region, 500 points were sampled uniformly at random. At each point, the MODIS Terra Surface Reflectance 8-Day time series was exported for January 1, 2018 to December 31, 2018 (Figure A12). MODIS collects 7 bands and NDVI was computed as an eighth feature, resulting in a time series of dimension  $8 \times 46$ . Global land cover labels came from the MODIS Terra+Aqua Combined Land Cover Product, which classifies every 500m-by-500m pixel into one of 17 land cover classes (e.g., grassland, cropland, desert).

**Comparison with Related Works** This SUSTAINBENCH dataset from [104] is the first time that any few-shot learning dataset has been released for satellite data. Because land cover products are available globally (albeit with varying accuracy), Wang et al. 104 created a few-shot dataset for land cover classification.

**Dataset Impact** Our hope is that this dataset can be included in evaluations of few-shot learning algorithms to see how they do on real-world time series, and that new algorithms will improve knowledge sharing from high-income regions to low-income ones. That way, performance on remote sensing tasks can be increased in low-income regions for tasks with few labels.

## E Benchmark Details

Code to reproduce baseline models new to SustainBench can be found in our GitHub repo.

### E.1 DHS survey-based regression tasks (SDGs 1, 3, 4, 6)

The DHS survey-based regression tasks include predicting an asset wealth index (SDG 1), women’s BMI and child mortality rates (SDG 3), women’s educational attainment (SDG 4), and water and sanitation indices (SDG 6). We adapt the KNN `scalar` NL model from [109] as the SUSTAINBENCH baseline model for these tasks. We chose this model for its simplicity and its high performance on predicting asset wealth as noted in [109]. For each label, we fitted a  $k$ -nearest neighbor ( $k$ -NN) regressor implemented using scikit-learn, and the  $k$  hyperparameter was tuned on the validation split, taking on integer values between 1 and 20, inclusive. The input to the  $k$ -NN model is the mean nightlights value from the nightlights band in the satellite input image, with separate models trained for the DMSP (survey year  $\leq 2011$ ) vs. VIIRS (survey year  $\geq 2012$ ) bands.

**Comparison with Related Works** We observe that our KNN nightlights baseline model roughly matches the performance described in [109] on the poverty prediction over space task ( $r^2 = 0.63$ ). However, its  $r^2$  values for predicting the other non-poverty labels is much lower: child mortality rate ( $r^2 = 0.01$ ), women BMI (0.42), women education (0.26), water index (0.40), sanitation index (0.36). Our result is in line with a similar observation made by [45], which also found that models trained on satellite images were better at predicting the asset wealth index than other non-poverty labels in 4 African countries. This strongly suggests that predicting these other labels almost certainly requires different models and/or inputs. Indeed, this is why SUSTAINBENCH provides street-level imagery in addition to satellite imagery.

While SUSTAINBENCH also provides street-level images for many DHS clusters, we do not have any baseline models yet that take advantage of the street-level imagery. Some preliminary results using street-level imagery to predict asset wealth and women’s BMI are shown in [64], although they only tested their models on India and Kenya (compared to the  $\sim 50$  countries included for DHS-based tasks in SUSTAINBENCH). We encourage researchers to develop new methods that can utilize both satellite imagery and street-level imagery, where available.

### E.2 SDG 2: Zero Hunger

#### E.2.1 Cropland mapping

Following Wang et al. 102, this task evaluates the model’s performance on semantic segmentation. The goal for the task with a single pixel label is to predict whether the single labeled pixel in the image is cropland or not. The goal for the task with image-level labels is to detect whether the majority ( $\geq 50\%$ ) of pixels in an image are classified to the cropland category. In both cases, the model is a U-Net trained using the binary cross entropy loss defined as

$$l(y, \hat{y}) = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})], \tag{1}$$

where  $y$  is either the single-pixel label or the image-level binary label and  $\hat{y}$  is the single-pixel or image-level model prediction. The evaluation metric is test set accuracy, precision, recall, and F1 scores. Details about the dataset are provided in Appendix D.3.

**Comparison with Related Works** As mentioned in Appendix D.3, existing cropland products have been created using SVMs or tree-based algorithms that take into account a single pixel at a time [18, 35, 106]. In Togo, Kerner et al. 56 showed that a multi-headed LSTM (still trained on single pixels) outperformed these classifiers on cropland prediction. Since SUSTAINBENCH’s cropland dataset is a static mosaic over the growing season, we chose to stick with the U-Net in Wang et al. 102 as the backbone architecture for the baseline. Segmentation models that are more state-of-the-art

than the U-Net would be good candidates to surpass this baseline. Active learning or semi-supervised learning methods could also beat a baseline that uses randomly sampled weak labels for supervision. Future updates to this cropland dataset can include the temporal dimension for cropland mapping as well.

### E.2.2 Crop type mapping in Ghana and South Sudan

The architecture described in Rustowicz et al. 83 obtained an average F1 score and overall accuracy of 0.57 and 0.61 in Ghana and 0.70 and 0.85 in South Sudan respectively, demonstrating the difficulty of this task. We use the same train, validation, and test splits as [83]. However, we use the full  $64 \times 64$  imagery provided, while [83] further subdivided imagery into  $32 \times 32$  pixel grids due to memory constraints. We also include variable-length time series with zero padding and masking, while [83] trimmed the respective time series down to the same length. We include variable-length time series with the reasoning that future research should be extendable to variable length time-series imagery. Due to these changes, we do not include baseline models from [83] for this iteration of the dataset. We provide more details in Appendix D.4.

**Comparison with Related Works** Like cropland maps, most operational works classifying crop types employ SVM or random forest classifiers [69, 49]. The baseline model that we use from Rustowicz et al. 83 improves upon these by using an LSTM-CNN. Recent models used in other, non-operational works include 1D CNNs and 3D CNNs [103] and kNN [55]. A review from this year comparing five deep learning models found that 1D CNN, LSTM-CNN, and GRU-CNN all achieved high accuracy on classifying crop types in China, with differences between them statistically insignificant [111].

### E.2.3 Crop type mapping in Kenya

The crop type data in Kenya come from three regions: Bungoma, Busia, and Siaya. We provide ML researchers with the option of splitting fields randomly or by region. The former setup would test the crop type classifier’s ability to distinguish crop type in-domain, while the latter would test the classifier’s out-of-domain generalization. In Table 2, we show results for the latter from [58].

In Kluger et al. 58, the authors trained on one region and tested on the other two in order to design algorithms that transfer from one region to another. In order to generalize across regions, they corrected for (1) crop type class distribution shifts and (2) feature shift between regions by estimating the shift using a linear model. The features used are the coefficients of a harmonic regression on Sentinel-2 time series. (In the field of remote sensing, the Fourier transform is a common way to extract features from time series [54].) The results from Kluger et al. 58 show that harmonic features achieve a macro F1-score of 0.30 when averaged across the three test sets, highlighting the difficulty of this problem. Note that this baseline did not include the Non-crop class in the analysis.

**Comparison with Related Works** We expect that, for in-domain crop type classification, methods mentioned previously (1D CNN, LSTM-CNN, GRU-CNN) will outperform the random forests and LDA used in [54] and [58]. However, for cross-region crop type classification, Kluger et al. 58 found that a simpler LDA classifier outperformed a more complex random forest. Nonetheless, deep learning-based algorithms that are designed for out-of-domain generalization could outperform the baseline. To our knowledge, these methods have not yet been tested on crop type mapping.

### E.2.4 Crop yield prediction

The task is to predict the county-level crop yield for that season, in metric tonnes per cultivated hectare, from the MODIS spectral histograms. We split the task into three separate subtasks of crop yield prediction in the United States, Argentina, and Brazil, and provide a 60-20-20 train-validation-test split. For each subtask, we encourage the usage of transfer learning and other cross-dataset training, especially due to the imbalance in data availability, between the United States, Argentina, and Brazil. Averaged across the years from 2012–2016, the benchmark models in Wang et al. 101 achieve an RMSE of 0.62 trained and evaluated on Argentina, 0.42 trained and evaluated on Brazil, and 0.38 using transfer learning on an Argentina-trained model to evaluate on Brazil. Averaged across 2011–2015, the benchmark models in You et al. 110 achieve an RMSE of 0.37 trained and evaluated

on the United States. However, we note that our datasets and splits are not identical to the original papers, so the results are not directly transferable.

**Comparison with Related Works** Several past works apply machine learning algorithms to human-engineered satellite features such as linear regression over NDVI [76] and EVI2 [16]. The papers that originally compiled SUSTAINBENCH’s datasets compared against these methods and outperformed them. A few other works, like Sun et al. 90, apply different architectures to spectral histograms similar to those provided in SUSTAINBENCH. Still other methods report results trained on ground-based data, such as ground-level images of crops [91], but these datasets have not been made public.

### E.2.5 Farmland parcel delineation

Given an input satellite image, the goal is to output the delineated boundaries between farm parcels, or the segmentation masks of farm parcels [9]. Similar to [9], given the predicted delineated boundaries of an image, we use the Dice score as the evaluation metric

$$DICE = \frac{2TP}{2TP + FP + FN}, \quad (2)$$

where “TP” denotes True Positive, “FP” denotes False Positive, and “FN” denotes False Negative. As discussed in [9], the Dice score Equation (2) has been widely used in image segmentation tasks and is often argued to be a better metric than accuracy when class imbalance between boundary and non-boundary pixels exists.

**Comparison with Related Works** 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. To our knowledge, this is the state-of-the-art deep learning model for field delineation.

## E.3 SDG 13: Climate Action

The task is binary classification on satellite imagery, where class 0 "no kiln" means there is no brick kiln present in the image and class 1 "yes kiln" means there is a brick kiln. The training-validation split of the provided Sentinel-2 imagery is 80-20. The ResNet50 [43] model trained in [63] achieved 94.2% accuracy on classifying high-resolution (1m/pixel) imagery; the authors hand-validated all positive predictions and 25% of negative predictions. The imagery was not released publicly because it was proprietary, so we report a baseline validation accuracy of 94.5%, training a ResNet50 model on lower-res Sentinel-2 imagery using only the Red, Blue, and Green bands (B4, B3, B2). In addition to accuracy on the validation set, AUC, precision, and recall are also valuable metrics given the class skew toward negative examples.

## E.4 SDG 15: Life on Land

### E.4.1 Representation learning for land cover classification

Jean et al. 53 performed land cover classification using features learned through an unsupervised, contrastive loss algorithm named Tile2Vec. Since the features are learned in entirely unsupervised ways, they can be used with any number of labels to train a classifier. At  $n = 1000$ , Tile2Vec features with a multi-layer perceptron (MLP) classifier achieved 0.55 accuracy; at  $n = 10,000$ , Tile2Vec features with an MLP achieved 0.58 accuracy. Notable also is that Tile2Vec features outperformed end-to-end training with a CNN sharing the same architecture as the feature encoder up to  $n = 50,000$  labels.

**Comparison with Related Works** Jean et al. 53 was the first to apply the distributional hypothesis from NLP to satellite imagery in order to learn features in an unsupervised way. Tile2Vec features outperformed features learned via other unsupervised algorithms like autoencoders and PCA. Methods that have not yet been tried but could yield high-quality representations include inpainting missing tiles, solving a jigsaw puzzle of scrambled satellite tiles, colorization, and other self-supervised learning techniques. Recently, [80] proposed a representation learning approach that uses randomly sampled patches from satellite imagery as convolutional filters in a CNN encoder, which could also be tested on this dataset.

## E.4.2 Out-of-domain land cover classification

Wang et al. 104 defined 1-shot, 2-way land cover classification tasks in each region, and compared the performance of a meta-learned CNN with pre-training/fine-tuning and training from scratch. The meta-learned CNN performed the best on the meta-test set. The meta-learning algorithm used was model-agnostic meta-learning (MAML). The MAML-trained model achieved an accuracy of 0.74, F1-score of 0.72, and kappa score of 0.32 when averaged over all regions in Sub-Saharan Africa in the meta-test set. Unlike other classification benchmarks in SUSTAINBENCH, this benchmark uses the kappa statistic to evaluate models because accuracy and F1-scores can vary widely across regions depending on the class distribution, and it is not clear whether an accuracy or F1-score is good or bad from the values alone.

We note that, as previously mentioned, existing land cover products tend to be less accurate in low-income regions such as Sub-Saharan Africa than in high-income regions. As a result, the MODIS land cover product used as ground truth will have errors in low-income regions. We suggest users also apply meta-learning and other transfer learning algorithms using other continents (*e.g.*, North America, Europe) as the meta-test set for algorithm evaluation purposes.

**Comparison with Related Works** To our knowledge, [104] and [82] (same authors) were the first works to apply meta-learning to land cover classification in order to simulate sharing knowledge from high-income regions to low-income ones. The baseline cited in Table 2 uses MAML, which is one of the most widely-used meta-learning algorithms. As the field of meta-learning is advancing quickly, we hope ML researchers will evaluate the latest meta-learning algorithms on this land cover classification dataset.

## F Ethical Concerns

Because the SDGs are high stakes issues with direct societal impacts ranging from local to global levels, it is imperative to exercise caution in addressing them. Researchers must be aware of and work to address the potential biases in the training data and in the generated predictions. For example, current models have been observed to over-predict wealth in poor regions and under-predict wealth in rich regions [52]. If such a model were used to distribute aid, the poor would receive less than they should. Much work remains to be done to understand and rectify the biases present in ML model predictions before they can play a significant role in policy-making.

Because the SUSTAINBENCH dataset involves remote sensing and geospatial data that covers areas with private property, data privacy can be a concern. We summarize below the risks of revealing information about individuals present in each dataset.

- For our survey data (see Tables A3 and A6), the geocoordinates for DHS and LSMS survey data are jittered randomly up to 2km for urban clusters and 10km for rural clusters to protect survey participant privacy [19]. Furthermore, geocoordinates and labels are only released for “clusters” (roughly villages or small towns); no household or individually identifiable data is released.
- Mapillary images, as well as satellite images from Landsat, Sentinel-1, Sentinel-2, MODIS, DMSP, NAIP, and PlanetScope, are all publicly available. In particular, all of these satellites other than PlanetScope are low-resolution. Mapillary automatically blurs faces of human subjects and license plates, it allows users who upload images to manually blur parts of images for privacy. Thus it is very difficult to get individually identifiable information from these images, and we believe that they do not directly constitute a privacy concern.
- The crop yield statistics, made publicly available by the governments of the US, Argentina, and Brazil, are published after aggregating over such large areas that the yields of individual farms cannot be derived.
- The crop type dataset released by Rustowicz et al. 83 has no geolocation information that would allow tracing to individuals. The satellite imagery released also has noise added so that it is more difficult to identify the original location and time that the imagery was taken. The crop type dataset released in Kenya likewise does not include geolocation.

- For the field delineation dataset, boundary shapefiles are publicly available from the French government as part of the European Union’s Common Agricultural Policy [9]. The data has been stripped of any identifying information about farmers.
- Brick kilns labels were generated by one of the authors under the guidance of domain experts. The version of this dataset released in SUSTAINBENCH consists of Sentinel-2 imagery, from which very few privacy-concerning details can be seen (see Figure A10).
- The labels used for the representation learning task and out-of-domain land cover classification task are products of other machine learning algorithms. They are publicly available and do not reveal information about individuals.