KidSat: satellite imagery to map childhood poverty dataset and benchmark

CHILDREN'S LIVES: INTERNATIONAL CONFERENCE ON CHILDREN AND THEIR FAMILIES USING THE MULTIPLE INDICATOR CLUSTER SURVEYS (MICS)

Makkunda Sharma, Fan Yang, Duy-Nhat Vo, Esra Suel, Swapnil Mishra, Samir Bhatt, Oliver Fiala, William Rudgard, Seth Flaxman

Our Contributions :

- Proposed a new dataset pairing satellite imagery and survey data on child poverty
- Created a benchmark task to measure childhood severe deprivation based on satellite image features
- Implemented baselines for the task, ranging from foundation vision models to satellite imagery based models
- Released the dataset downloading, benchmark creation and baseline code

https://github.com/MLGlobalHealth/KidSat

Dataset

- **Satellite Images** - 33,608 images, each 10 km × 10 km based on DHS survey cluster locations
- **Image Source** - Landsat 5,7 and 8, and Sentinel-2 (from 2015 onwards)
- **Years** - 1997 to 2022
- **Countries** - 19 countries from Eastern and Southern Africa
- **Survey** - Cluster level geolocated DHS survey

Benchmark

Severe Deprivation:

As per UNICEF definition, a child is considered severely deprived if they experience severe deprivation in at least one of the following areas :

- **Housing**: Severe deprivation if the number of persons per room is 5 or more
- **Water**: Severe deprivation if the household uses unsafe water sources
- **Sanitation**: Severe deprivation if the household lacks access to safe sanitation facilities
- **Nutrition**: Severe deprivation if a child's height-for-age z-score is below -3 (indicating severe stunting)
- **Health**: Severe deprivation if a child misses all essential vaccines or has untreated acute respiratory infections
- **Education**: Severe deprivation if a school-aged child does not attend school and has not completed any level of education

Table 2: DHS variables selected for model fine-tuning

Variable	Description	Type
h10	Whether the child ever received any vaccination to prevent diseases.	Categorical
h3	DPT 1 vaccination.	Categorical
h31	Whether the child had suffered from a cough in the last two weeks and whether the child had been ill with the cough in the last 24 hours.	Categorical
h5	DPT 2 vaccination.	Categorical
h7	DPT 3 vaccination.	Categorical
h9	Measles 1 vaccination.	Categorical
hc70	Height for age standard deviation (according to WHO).	Continuous
hv106	Highest level of education the household member attended.	Categorical
hv109	Educational attainment recoded.	Categorical
hv121	Household member attended school during current school year.	Categorical
hv201	Main source of drinking water for members of the household.	Categorical
hv204	Time taken to get to the water source for drinking water	Continuous
hv205	Type of toilet facility in the household.	Categorical
hv216	Number of rooms used for sleeping in the household.	Continuous
hv225	Whether the household shares a toilet with other households.	Categorical
hv271	Wealth index factor score (5 decimals)	Continuous
v312	Current contraceptive method.	Categorical

Source : https://arxiv.org/pdf/2407.05986

Tasks:

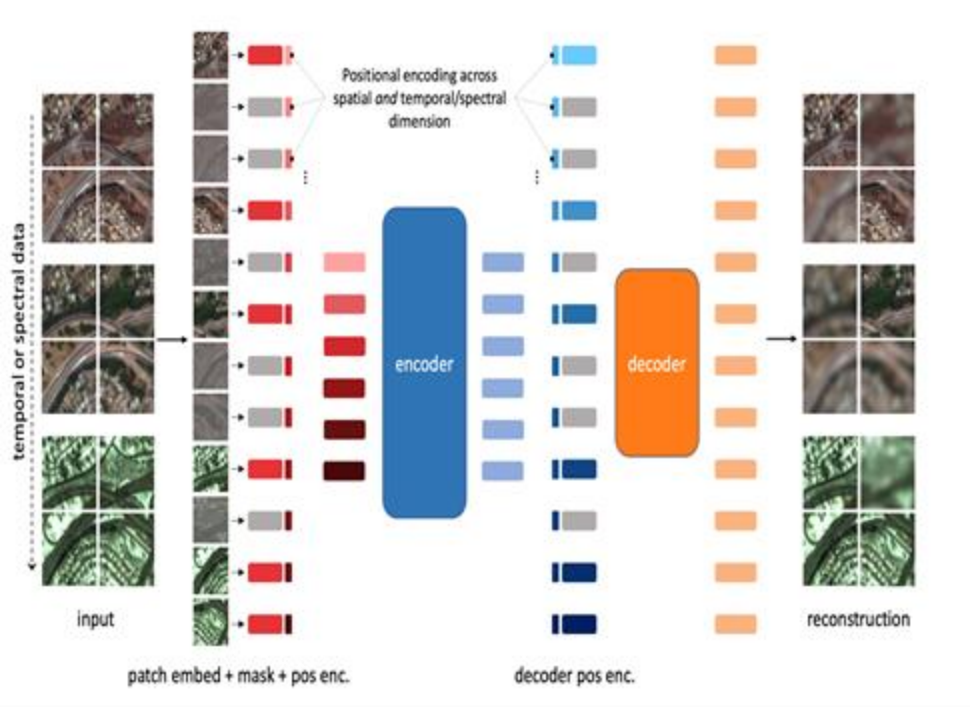
- **Spatial :**
  - We would like to test generalization across locations
  - 5-fold spatial cross-validation at the cluster level across the dataset.
  - Training dataset is 80% of the clusters, and performance is tested on the held out 20%.
- **Temporal :**
  - We would like to test generalization across time
  - Training data for this task is the dataset from 1997 to 2019
  - Performance is tested on the data from 2020 to 2022

Prediction

Baselines :

- **Mean Prediction :** The simplest baseline, where we predict training data mean at every location in the test set.
- **Gaussian Process Regression** : We run a gaussian process regression, where the input is the latitude and longitude of the cluster, and the target is the severe deprivation variable at the location
- **MOSAIKS** :
  - It is generalisable feature extraction framework for satellite images developed for environmental and socio-economic applications
  - We use the MOSAIKS api to get a feature vector for the sentinel-2 satellite image 10 km x 10 km at the cluster location
  - We run a ridge regression from the feature vector to the target variable

SatMAE



- SatMAE, a pre-training framework for temporal or multi-spectral satellite imagery based on Masked Autoencoder (MAE).
- The base SatMAE model is the decoder of a MAE, outputting a 1024-dimensional vector.
- Fine-tuning is done by appending a transformer head mapping to the 99-dimensional target DHS variable created from the survey .
- Finally, we perform ridge regression from the 99 dimensional vector to the target variable

Source : https://arxiv.org/pdf/2207.08051

DINOv2

- DINOv2 is a general purpose foundation vision model, trained by unsupervised learning on a corpus of hundreds of millions of images
- The underlying model is a base ViT (Vision Transformer) 1B (1 billion parameters)
- The model outputs a 768 dimension feature vector
- For fine tuning we appended a regression head to predict the 99 dimension DHS vector
- For prediction, we perform ridge regression from the 99 dimensional vector to the target variable
- Fine tuned DINOv2 was the best performing model on our dataset on both tasks

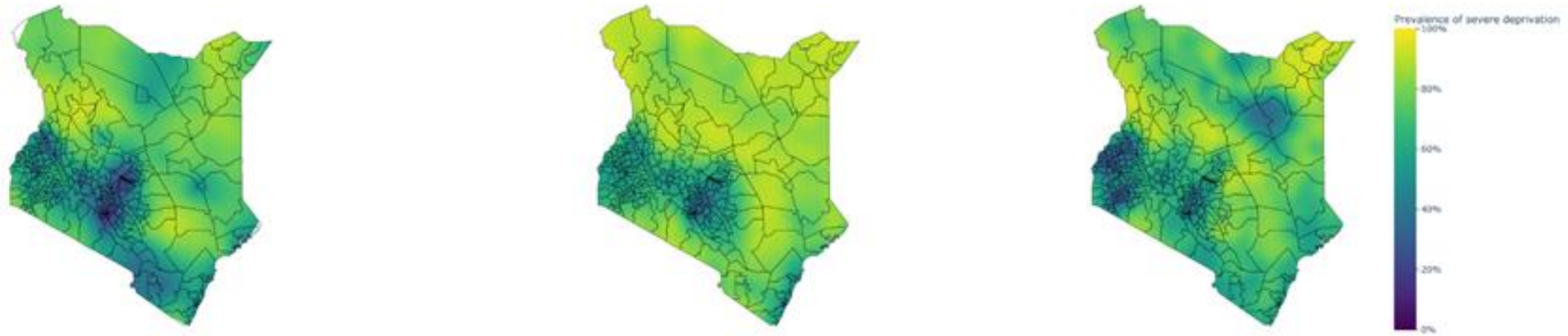
Results

Model	Benchmark Type	MAE ± SE (Spatial)	MAE (Temporal)
Mean Prediction	-	0.2930 ± 0.0018	0.3183
Gaussian Process Regression	-	0.2436 ± 0.0002	0.5656
MOSAIKS	Sentinel-2	0.2356 ± 0.0114	0.2588
DINOv2 (Raw)	LandSat	0.2260 ± 0.0005	0.2704
DINOv2 (Raw)	Sentinel-2	0.2013 ± 0.0019	0.2597
DINOv2 (Fine-tuned)	LandSat	0.2042 ± 0.0015	0.2574
DINOv2 (Fine-tuned)	Sentinel-2	0.1836 ± 0.0036	0.2858
SatMAE (Raw)	LandSat	0.2341 ± 0.0017	0.3453
SatMAE (Raw)	Sentinel-2	0.2347 ± 0.0027	0.3067
SatMAE (Fine-tuned)	LandSat	0.2125 ± 0.0019	0.3376
SatMAE (Fine-tuned)	Sentinel-2	0.2093 ± 0.0039	0.3139

Source : https://arxiv.org/pdf/2407.05986

Applications

DHS - Kenya 22:



(a) Kriging estimates using Kenya 2022 DHS

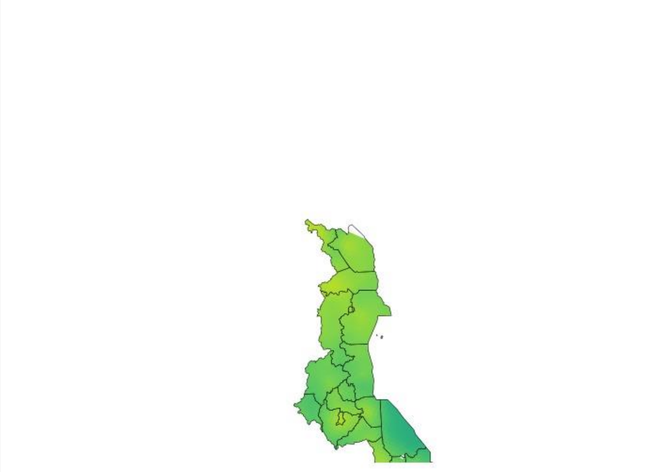
(b) DINOv2 fine-tuned on KidSat spatial training dataset

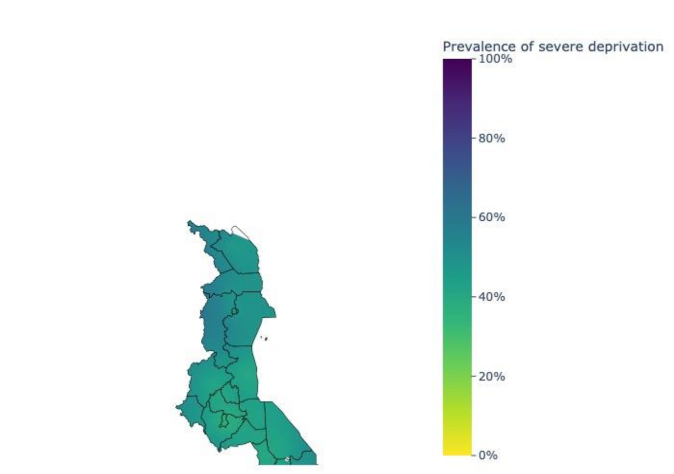
(c) DINOv2 fine-tuned on KidSat temporal training dataset

Estimates of the prevalence of severe deprivation for Kenya in 2022

Source : https://arxiv.org/pdf/2407.05986

MICS – Malawi 19-20:





Estimates of the prevalence of severe deprivation in Malawi

Extensions and Future Work

Possible Extensions:

- We would like to evaluate the performance of various models on the six components of child poverty separately, on moderate as opposed to severe deprivation.
- We evaluated spatial generalisation by leaving out clusters, a stricter evaluation would have considered a leave-one-country out evaluation.
- Instead of just evaluating on MICS data, finetuning the models, and comparing performance to the DHS survey

Acknowledgements:

- Alan Turing Institute
- EPSRC
- National Research Foundation
- MRC Centre for Global Infectious Disease Analysis
- National Institute for Health and Care Research (NIHR)
- Novo Nordisk Foundation
- Danish National Research Foundation
- The Eric and Wendy Schmidt Fund For Strategic Innovation
- Google Cloud for Researchers

Thank you