Time Series of Satellite Imagery Improve Deep Learning Estimates of Neighborhood-Level Poverty in Africa

Citation for the original published paper (version of record):
Pettersson, M., Kakooei, M., Ortheden, J. et al (2023). Time Series of Satellite Imagery Improve
Deep Learning Estimates of Neighborhood-Level Poverty
in Africa. IJCAI International Joint Conference on Artificial Intelligence, 2023-August: 6165-6173.
http://dx.doi.org/10.24963/ijcai.2023/684

(article starts on next page)

Time Series of Satellite Imagery Improve Deep Learning Estimates of Neighborhood-Level Poverty in Africa

Markus B. Pettersson[1,4*] , **Mohammad Kakooei**[1,4] , **Julia Ortheden**[1] ,
Fredrik D. Johansson[1] and **Adel Daoud**[1,2,3,4]
[1]Data Science and AI Division, Chalmers University of Technology, Sweden
[2]Institute for Analytical Sociology, Linköping University, Sweden
[3]Center for Advanced Study in the Behavioral Sciences, Stanford University, United States
[4]The AI and Global Development Lab
{markus.pettersson, kakooei, fredrik.johansson, daoud}@chalmers.se

Abstract

To combat poor health and living conditions, policymakers in Africa require temporally and geographically granular data measuring economic well-being. Machine learning (ML) offers a promising alternative to expensive and time-consuming survey measurements by training models to predict economic conditions from freely available satellite imagery. However, previous efforts have failed to utilize the temporal information available in earth observation (EO) data, which may capture developments important to standards of living. In this work, we develop an EO-ML method for inferring neighborhood-level material-asset wealth using multi-temporal imagery and recurrent convolutional neural networks.[1] Our model outperforms state-of-the-art models in several aspects of generalization, explaining 72% of the variance in wealth across held-out countries and 75% held-out time spans. Using our geographically and temporally aware models, we created spatio-temporal material-asset data maps covering the entire continent of Africa from 1990 to 2019, making our data product the largest dataset of its kind. We showcase these results by analyzing which neighborhoods are likely to escape poverty by the year 2030, which is the deadline for when the Sustainable Development Goals (SDG) are evaluated.

1 Introduction

About 700 million people live in extreme poverty, i.e., on less than $2.15 a day, and about two-thirds of these people live in Africa [World Bank, 2022]. Although researchers have a firm understanding of poverty at the level of countries [Halleröd *et al.*, 2013], granular geo-temporal data is necessary to effectively target social and economic policy at the neighborhood level. Classically, researchers and policymakers have collected such data through surveys [Daoud and Dubhashi, 2023; Groves and Lyberg, 2010; Lavrakas, 2008], but these are expensive to perform, are geographically limited, and seldom provide information about the historical trajectory of poverty which may influence policy.

Consequently, a major research challenge is how best to fill in the gaps in the geo-temporal poverty data for neighborhoods and moments where no surveys have been conducted. Unlike surveys, planetary-scale satellite imagery is available in Africa over both wide and granular geography and temporality [Gorelick *et al.*, 2017]. Researchers have shown that such earth observation (EO) data and machine learning (ML) models can be successfully used to impute proxies of poverty in Africa & India [Jean *et al.*, 2016; Yeh *et al.*, 2020; Chi *et al.*, 2022; Rolf *et al.*, 2021; Daoud *et al.*, 2023].

While EO-ML models represent a significant breakthrough in poverty research, they provide only a granular geographical "snapshot" of African poverty—they fail to leverage and estimate its temporality. From earth's orbit, it is possible to observe the economic development of neighborhoods through the construction and maintenance of housing, farmland and infrastructure such as roads, bridges and factories. Such progress is directly associated with the wealth of the people living in the neighborhood. Conversely, if an area is not expanding with time, this can indicate a lack of resources [Daoud, 2011]. A temporally aware model, which observes this development, or lack thereof, can better estimate levels of and changes in poverty.

In this article, we develop a new EO-ML model for imputing poverty levels from satellite imagery which is both geographically and temporally aware. To do this, we assemble 138 African surveys stretching over a 30-year period from 1990-2020, containing data on the health and living conditions of household clusters, providing the labels on economic status needed for training. We map each survey point to its corresponding raw satellite images, providing the input data to our model. Second, we generalize an existing EO-ML architecture proposed by Yeh et al. [2020], based on a *residual neural network* (ResNet), to accept sequences of satellite images as input using a long short-term memory (LSTM) layer [Hochreiter and Schmidhuber, 1997]. This allows the model to learn from both geographical and temporal features.

Our experiments show that our temporal EO-ML method

[*]Contact author

[1]Technical appendix and code can be found at the project page: github.com/AIandGlobalDevelopmentLab/temporal-eo-wealth

increases Pearson's correlation (r^2) by up to 29.3% on held-out data. Based on our geo-temporal EO-ML model, we create poverty estimates for every populated neighborhood location for the whole continent of Africa. To the best of our knowledge, this is the largest estimated poverty data of its kind that covers African neighborhood-level poverty at this level of granularity (6.72 × 6.72 km) and over 30 years [Chi *et al.*, 2022]. Our results correlate strongly with country-level economic statistics and trends. Finally, we forecast which neighborhoods are likely to escape poverty by the year 2030—the year in which the United Nations' Sustainable Development Goals (SDGs) will be evaluated. This use case exemplifies how our data can be used to study planetary problems on a different scale than what has previously been possible using only earth observation data and machine learning.

2 Methodology

Our problem is to predict the average material wealth $Y_{i,t} \in \mathbb{R}$ in a cluster of households surrounding a location i at time t. The cluster is the primary sampling unit of DHS surveys, and it corresponds to a neighborhood in an urban setting and a village in rural areas. We use as input a sequence of satellite imagery $\boldsymbol{X}_{i,t} \in \mathbb{R}^{F \times w \times h \times d}$ covering the location i and surrounding the time point t. Here, F represents the number of image frames, w and h the image width and height in pixels, and d the number of bands. Each one of the F image frames is equally spaced out in time such that they each represent a time span of the same fixed length. This paper aims to develop an EO-ML method which given $\boldsymbol{X}_{i,t}$ predicts $Y_{i,t}$.

2.1 Data

Survey Data

In this article, we use data from ∼1.2 million households living in 57,195 survey clusters in 36 African countries. The data was drawn from 138 nationally representative Demographic and Health Surveys (DHS) conducted between 1991 and 2019. The DHS program has conducted surveys with standardized questions in low- to middle-income countries, and since the 1990's many of these surveys contain GPS coordinates [DHS, 2022]. Although the questionnaires are aimed at collecting household-level information, the corresponding geo-location is collected for the centroid of a neighborhood to which a household belongs. To preserve household privacy, the DHS displaces the GPS coordinate by up to 2 km for urban neighborhoods and up to 5 km for 99% for rural locations, with the remaining 1% of given coordinates displaced at most 10 km from the true location [Burgert *et al.*, 2013]. Because these surveys are cross-sectional, we only measure each neighborhood once.

To obtain a standardized indicator of cluster-level asset wealth Y, we used the International Wealth Index (IWI), developed by Smits et al. [2015]. The IWI variable is computed from household-level answers to the DHS questionnaire asking whether members of the household have access to the following: TV, refrigerator, phone, bike, car, utensils, and electricity. The questionnaire also measures the quality of facilities: water, toilet, floor, and the number of bedrooms. The IWI is then obtained by taking the first principal component of the questionnaire responses. The purpose of the IWI is to function as a one-dimensional summary of the many dimensions of human health and living conditions. After each household's IWI has been estimated, we calculate the mean IWI for each neighborhood (cluster), such that each cluster i surveyed at time t receives a corresponding mean IWI score, $Y_{i,t} \in [0, 100]$. We obtained the IWI by using the R package DHSharmonisation [Ekbrand, 2019].

Satellite Imagery

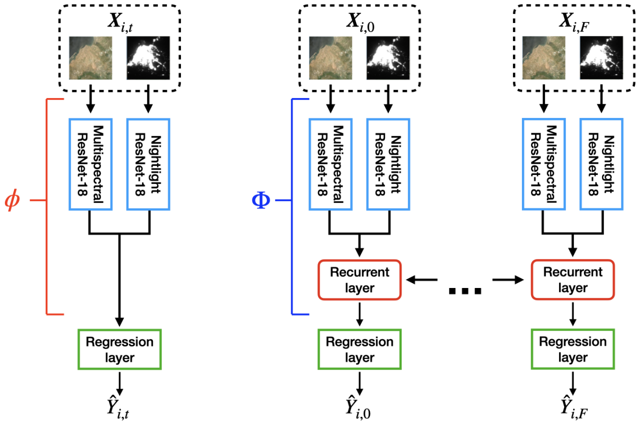
Our pipeline for gathering and preprocessing EO data largely follows that of Yeh et al. [2020]. The main source of EO data were surface reflectance images from the Landsat 5, 7, and 8 satellites [Gorelick *et al.*, 2017]. To mitigate seasonal variations, input frames were compiled from the per-band three-year median of all cloud-free pixels. Since our study spans the 30-year period from 1990 to 2019, the complete Landsat dataset was condensed into ten sequential image-composite frames. For every cluster (i), we collected a sequence of ten frames centered on the corresponding survey location, forming a concise video of the area measuring 6.72 × 6.72 km. Further details about the chosen image size can be found in Appendix A.1; derived from our CNN architecture's input size (224 × 224 pixels) and the Landsat image resolution (30 m/px). We used all seven available Landsat bands, which we refer to as *multispectral*: red, green, blue, near-infrared, shortwave infrared 1, shortwave infrared 2, and thermal.

We included nighttime luminosity to these seven bands, as previous works have shown that it is of economic activity [Henderson *et al.*, 2012]. We used images from the DMSP [Hsu *et al.*, 2015] satellite for the first seven frames (1990-2010) and the VIIRS [Elvidge *et al.*, 2017] satellite for the final three (2011-2019). As nightlights have a lower resolution (1 km/px for DMSP and 450 m/px for VIIRS) than the Landsat daylight images (30 m/px), they were resized using nearest-neighbor upsampling to cover the same spatial area as the Landsat images. They were then appended as an extra band to the composite frames. All EO data was processed using the Google Earth Engine [Gorelick *et al.*, 2017].

2.2 Method

A SINGLE-FRAME Model: The Baseline Architecture

Yeh et al. [2020] proposed a convolutional neural network (CNN) architecture for learning to predict asset wealth, as measured in national wealth surveys, based on multispectral daytime- and nighttime luminosity images captured by satellites. This model, illustrated in Figure 1a, contains two architectural segments. The first, the encoder $\phi : \mathbb{R}^{w \times h \times b} \to \mathbb{R}^{2v}$, separates the daylight and nighttime bands, using the well-known ResNet-18 architecture (v2 with pre-activation) [He *et al.*, 2016] to produce one feature representation for the daytime and one for nightlight imagery. Here, $w \times h \times b$ are the width, height and depth dimensions of a single image frame and v is the dimension of the last layer of a ResNet. In our case $v = 1000$. The encoding ϕ concatenates two v-dimensional encodings from two ResNets, one for the daytime and one for the nightlight bands. In the second segment, the pair of encodings are concatenated and combined with a linear layer $\hat{y} = \beta^\top \phi(\cdot)$, with parameters $\beta \in \mathbb{R}^{2v}$ to predict



(a) SINGLE-FRAME (b) MULTI-FRAME

Figure 1: The two architectures used for experiments. (a) is an identical architecture to the one proposed by Yeh et al. [2020]. It takes a single image, with multispectral and nightlight bands, as input. (b) is the same architecture, but with an added recurrent layer. Its input will consist of a time series of F multispectral-nightlight images.

wealth outcomes. This model, which we will refer to as the SINGLE-FRAME model, is trained to predict wealth from an image frame X_i centered on a surveyed cluster i by picking up on different spatial patterns correlated with wealth.

Since surveys are repeated cross-sectionally, i.e., new cluster locations are randomly drawn for each survey, it follows that each location i is only surveyed at a single time point t_i. Thus, we can simplify notation such that $y_i = Y_{i,t_i} \in \mathbb{R}$ represents available supervision for i. Similarly, there exists a single image frame X_i drawn from the multi-temporal sequence representative of this time point: $X_i = \boldsymbol{X}_{i,t_i} \in \mathbb{R}^{w \times h \times b}$. Our learning objective is the ridge-regularized mean square error (MSE) of the model output,

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta^\top \phi\left(X_i\right)\right)^2 + \lambda \sum_{j=1}^{2v} |\beta_j|^2 \tag{1}$$

where n is the number of samples (survey points), β are the regression coefficients and λ is a constant controlling regularization strength. As will be further discussed in Section 3.1, λ was tuned using cross-validation.

Two MULTI-FRAME Models

To study the advantage of additional information provided by multi-temporal EO data, we compare the SINGLE-FRAME model to an extended MULTI-FRAME architecture. By expanding the model input from a single image frame to a sequence of frames and inserting a recurrent long short-term memory (LSTM) layer we hope to make the model aware of temporal economic development around cluster i. Such development may include the construction of new infrastructure, urban expansion, environmental changes or other factors relevant to the area's level of wealth. This temporally aware MULTI-FRAME architecture builds on the SINGLE-FRAME model by adding a hidden recurrent layer.

Just as for the SINGLE-FRAME, the MULTI-FRAME architecture consists of two segments (see Figure 1b). The first, the encoder $\Phi : \mathbb{R}^{F \times w \times h \times b} \rightarrow \mathbb{R}^{F \times u}$, transforms a sequence of F image frames by i) encoding the daylight and nightlight bands of each image frame using ResNets, and ii) concatenating the two encodings for each frame, as in ϕ. Finally, in Φ, each frame encoding is linked through a bidirectional LSTM with output size $u = 32$, which results in a sequence of u-dimensional representations of length F. The second segment is, just like for the SINGLE-FRAME model, a single linear layer with parameters β used to predict the poverty label corresponding to each frame. Parameters of both encoders and prediction layers are shared between all frames.

The MULTI-FRAME architecture accepts a sequence of similarly configured images as input, essentially forming a short video sequence of the location. We denote any multi-temporal sequence of length F depicting cluster i and containing the time-point t_i by $\boldsymbol{X}_i^F \in \mathbb{R}^{F \times w \times h \times b}$. The model outputs a prediction for each of the F time frames, but as only one of these corresponds to the survey time t_i, when the label was measured, this is the only output considered by the loss function. We use f_i to denote the index in $\boldsymbol{X}_i^F$ corresponding to t_i. The MULTI-FRAME loss function then becomes

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta^\top \Phi_{f_i}\left(\boldsymbol{X}_i^F\right)\right)^2 + \lambda \sum_{j=1}^{u} |\beta_j|^2 \tag{2}$$

Note that the time-point t_i could be covered by any of the F time frames, from the first to the last, and that there are therefore up to F number of valid configurations of $\boldsymbol{X}_i^F$ (explained in detail in Appendix A.2). During training, it resamples a new sequence configuration uniformly for each epoch.

A Shallow Baseline Model

To obtain a reference point for the performance of our deep models (i.e., our SINGLE-FRAME and MULTI-FRAME CNN models), we use a shallow baseline: a basic k-nearest neighbor model (KNN). This model predicts IWI based on the k locations with the most similar nightlight-values. For each image frame, a histogram of nightlight luminosity levels was created. These were then used as inputs for the algorithm (KNN) where k was tuned by cross-validation.

3 Experimental Setup

The main aim of this paper is to evaluate whether a geographically and temporally aware EO-ML method improves the estimation of health and living conditions in Africa from satellite imagery. To this end, we conducted three experiments:

- **Out-of-area (OOA)**: The model is trained on one set of clusters and tested on another set, without stratification.
- **Out-of-country (OOC)**: The model is trained on clusters from one set of countries and evaluated on clusters from held-out countries. This is done to evaluate how well the model generalizes across borders.
- **Out-of-time-span (OOTS)**: The model is trained on surveys from one period of time and evaluated on surveys from a different time span. This is done to evaluate how well the model generalizes over time.

In all three experiments, we trained and evaluated our EO-ML models on data splits from five-fold cross-validation (CV). For the OOA experiment, the only restriction on sample splitting was that images in the training set could not geographically overlap (regardless of time) with the images in the test set. This restriction prevents models from making a prediction for a location in the test set by recognizing it from the training set. Appendix A.3 delineates how all clusters were divided into collections where, if all members of a collection were put in the same CV fold, this problem would be avoided. To ensure an equal number of clusters per fold, the largest remaining collection was sorted into the fold with the fewest samples until all collections had been assigned to a fold. The same procedure was carried out for the OOC experiment with each country treated as a collection. For the OOTS experiment, the folds consisted of the five consecutive time spans which resulted in the most equal number of clusters across all folds. These time spans were: 1991-2003, 2004-2008, 2009-2012, 2013-2015, and 2016-2019.

For each CV fold, we trained one shallow KNN-model, one SINGLE-FRAME-model, and two MULTI-FRAME models with different numbers of time steps. The first of these MULTI-FRAME models considered ten frames, our full 30-years time period for which we have image data when making predictions. The second one only considered a subset of five frames (15 years), selected to cover the observed time-point t_i. Further details about these two configurations can be found in Appendix A.2. The main reason for trying two different sequence lengths is to evaluate the impact of more temporal data.

3.1 Model Training and Hyperparameter Selection

Aside from the sample splitting procedure, the training pipeline was the same across all three experiments and all model architectures. First, all image bands were normalized to be between 0 and 1 based on the min- and max-values in the training data for each CV fold. As the two nightlights satellites used different pixel value ranges, these were normalized separately. During the training of the deep models, we applied image augmentation by random horizontal and vertical flips, as well as brightness and contrast shifts for the multispectral bands. All encoders (ResNet-18, see Section 2.2) were initialized with weights pre-trained on the ImageNet data set [Deng *et al.*, 2009]. As was proposed by Yeh et al. [2020], the weights in the first layer in our models (for the red, green, and blue bands) are the same as the ImageNet model. For the other bands, the mean weights of the red-green-blue bands were used for initialization. The remaining ResNet layers were kept as they were while we initialized the weights in the LSTM and regression layers randomly using Glorot uniform initialization [Glorot and Bengio, 2010].

We trained our models to minimize the MSE loss using the Adam optimizer [Kingma and Ba, 2014], with batch size 64. Each model used 300 epochs for training and tuning, finally selecting the model iteration with the best loss on the validation set (early stopping). To find a suitable learning rate faster, it was tuned using population-based tuning [Li *et al.*, 2019]: 12 agents sampled a learning rate from a log-uniform distribution (between 10^{-10} and 10^{-1}). In every third epoch, the worst-performing quarter of the models was perturbed. These perturbed models were paired with one of the models from the best-performing quarter and copied their weights and parameters. Their learning rates were then perturbed again by a multiplicative factor of 0.8 or 1.2, except for in 25% of cases where they were re-sampled from the original log-uniform distribution.

The regularization parameter λ (weight decay) for the final regression layers was tuned separately after the population-based training was finished. This was done by freezing the weights of previous layers (ϕ and Φ, respectively), using them to encode the training set, and tuning the parameter with grid search. The tuning was conducted for each of the different experiment-specific groups in the training and validation set: for the OOA experiment, the tuning was done over each CV fold; for OOC, it was done over each country; and, for OOTS it was done for each year.

3.2 Evaluation Metrics

To evaluate the discriminative predictive power of trained models, we use the squared Pearson's correlation r^2 between the predicted IWI $\hat{y}_i$ and its held-out ground-truth value y_i. Pearson's r^2 captures correlations between predictions and labels without accounting for differences in scale. To measure quality in predictions on an absolute scale, we use the root mean squared error (RMSE). As RMSE is the standard deviation of predictive variation, it gives an absolute measure of how well our model is generalizing to previously unseen (held-out) data. The RMSE maintains the scale of the IWI score, which ranges between 0 and 100, and thereby, may be readily interpreted by domain experts. This interpretative analysis is facilitated by using an absolute index.

4 Results

Figure 2 show all aggregated results across the three experimental settings. The KNN baseline model achieves a moderate performance across all experiments, with r^2 ranging from 0.42 to 0.62. This confirms the literature finding that satellite images are predictive of IWI [Jean *et al.*, 2016; Yeh *et al.*, 2020; Chi *et al.*, 2022]. The three deep architectures achieve higher performance than baseline, across all experiments. Moreover, one of our key findings is that temporally aware models (FIVE-FRAME, TEN-FRAME) outperform models that are not explicitly temporally aware (SINGLE-FRAME, KNN). In other words, by injecting our EO-ML model with an LSTM layer, giving it the capacity to use multiple frames, our model improves its predictive performance. Compared to the SINGLE-FRAME model (i.e., the current state-of-the-art EO-ML method), our TEN-FRAME model improves predictive performance by the following percentage points: in OOA by 10, OOC by 12, and OOTS by 17. The absolute performance of the 10-frame model across the three experiments are: ($r^2 = 0.76$) for OOA, ($r^2 = 0.72$) for OOC, and ($r^2 = 0.75$) for OOTS. These results show that historical and future satellite images, across a neighborhood's location, contain sufficient signal for predicting IWI.

A second key finding is that, although the performance increase is slight, the TEN-FRAME model consistently outperforms the FIVE-FRAME model. This finding suggests that

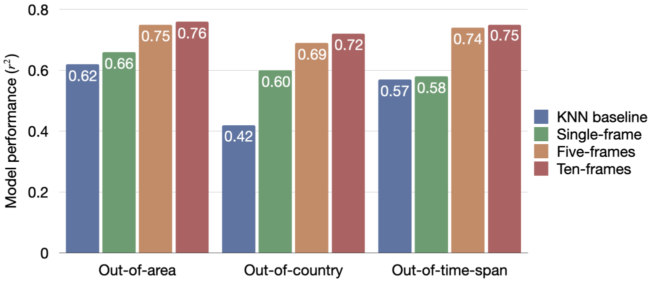


Figure 2: Performance of the baseline and the three model architectures trained on three different tasks. The MULTI-FRAME models are able to utilize temporal information (as described in Section 2.2) and thereby outperform the SINGLE-FRAME model across all three experiments.

the more frames used, the longer the period of time considered, and the better the predictions. Both findings reinforce our overarching hypothesis that by accounting for the inherent temporal structure in the data, an EO-ML methods will improve its performance, across both geography and temporality. In the following sections, we further probe our experimental results by analyzing models' performances by time, country, and locality (urban-rural).

4.1 Analyzing Results by Time

Because EO-ML models are beneficial for imputing IWI in time points that lack surveys, we assessed how well our models generalize to time spans they have not been exposed to during training. Figure 3 shows that both MULTI-FRAME models outperform the SINGLE-FRAME model across almost all years. On average, the FIVE-FRAME model improves the r^2 score by 0.14 above the SINGLE-FRAME, and the TEN-FRAME model improves by 0.13. The two MULTI-FRAME models exhibit a comparable performance for all years.

All models struggle with the starting decade, up to the year 2000. A possible explanation is that the year 1999 marks the launch of the Landsat 7 mission, which supplies more satellite images. All imagery prior to this is therefore solely captured by satellites from the Landsat 5 program, resulting in a domain shift for the input data from training to testing.

The SINGLE-FRAME model struggles in particular with the first year in which its r^2-performance approaches zero, whereas the MULTI-FRAME models hover just above $r^2 = 0.25$. Comparing the first and last year, we note that MULTI-FRAME performance is much higher in the ending frame. In addition, the best temporal performance occurs in the middle of the frame span, where r^2 approaches 0.9.

Figure 3 shows that model performances are highly variable across years. This variability has at least two explanations. First, it should be noted that it is impossible to fully differentiate between the model's performance in a given year from the model's performance in the countries surveyed that year. Because the DHS surveys only a subset of countries in any given year, we can only evaluate our models against these surveyed countries. Second, each year contains only a fraction of the countries for which we would like to evaluate predictions. For example, the years 1991 and 1996 only have surveys from a single country each, Cameroon and Benin, respectively; and in 2002, the DHS conducted no surveys at all. This lack of DHS surveys limits our training and evaluation capabilities, but it is an inherent data limitation. Nonetheless, despite these limitations, our MULTI-FRAME models demonstrate competitive average results, reaching $r^2 = 0.76$ for pooled observations across held-out folds (see Figure 2).

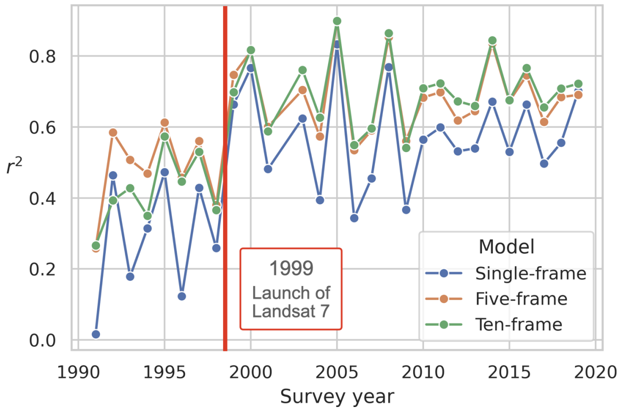


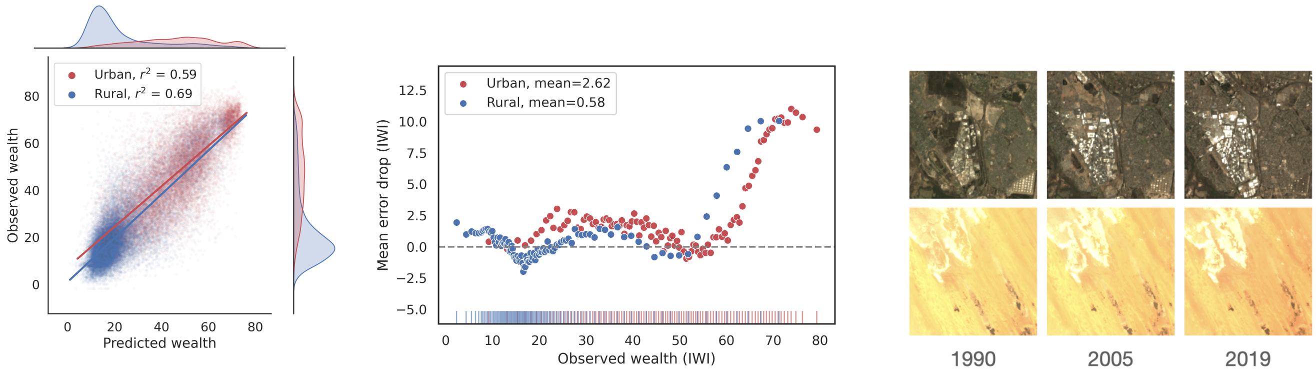
Figure 3: Performance of the OOTS models per year. In general, the MULTI-FRAME models outperform the SINGLE-FRAME. The models all struggle the most in the early years, with a big uptick in performance after the launch of Landsat 7.

4.2 Analyzing Results by Country

As one of our aims is to create a data map of the full continent of Africa, it is of particular interest to analyze the extent to which our EO-ML models generalize to countries not in the training set. That is the purpose of the OOC experiment.

For most countries, prediction performance is high (see Figure 9a in the appendix). The best-performing countries are diversely distributed, both geographically and socially, across the continent. The mean of country RMSEs of 8.82 (and a median of 8.12). As this figure is on the IWI scale (in the range $[0, 100]$), it provides an intuition for the distribution of predictive variation. The highest explained variation comes from Ethiopia where our FIVE-FRAME model trained on other countries can explain 80% of the variation in wealth between neighborhoods. For some countries, such as Morocco and South Africa, our model exhibits lower performance, with a minimum yet non-trivial $r^2 = 0.31$ for Lesotho.

In general, our MULTI-FRAME models generalize better to unseen countries than the SINGLE-FRAME model. As Figure 9b in the Appendix shows, for countries where the SINGLE-FRAME model is already performing well, the two MULTI-FRAME models improve performance only slightly. The largest improvements over the SINGLE-FRAME model can be found in countries where model performance was already poor. Thus, while our MULTI-FRAME models' performance is lowest in Lesotho, Morocco and South Africa, our models actually improve over the SINGLE-FRAME most in these particular. These countries have fewer samples, and are likely most different from the rest of the sample, making model learning hard in general.



(a) TEN-FRAME OOC model performance in urban versus rural regions

(b) Mean prediction improvement when switching from MULTI-FRAME to TEN-FRAME

(c) Examples of areas with a major gain versus deterioration in accuracy

Figure 4: Model performance in different localities. (a) Calibration of the FIVE-FRAME model in urban and rural locations. The model exhibits greater proficiency in distinguishing wealthier clusters from poorer ones in rural areas than in urban contexts. (b) Mean improvement when switching from the SINGLE-FRAME to the TEN-FRAME model aggregated for each percentile of surveyed IWI value and split by urban/rural. For both urban and rural settings, the largest improvements occur for relatively wealthy clusters. (c) Comparison between two locations with a big improvement/deterioration when switching from SINGLE-FRAME to MULTI-FRAME. The top row depicts an area with major developments where the SINGLE-FRAME model underestimated the level of wealth. The MULTI-FRAME models, which observed these changes, accurately gave a higher estimate.

4.3 Analyzing Results by Locality

To understand the performance of the models, we break down the results by urban and rural neighborhoods, using data available in DHS surveys. To reiterate, our EO-models do not have access to this information explicitly when trained. However, by adding locality to the predicted values, we are enabled to probe model performance better.

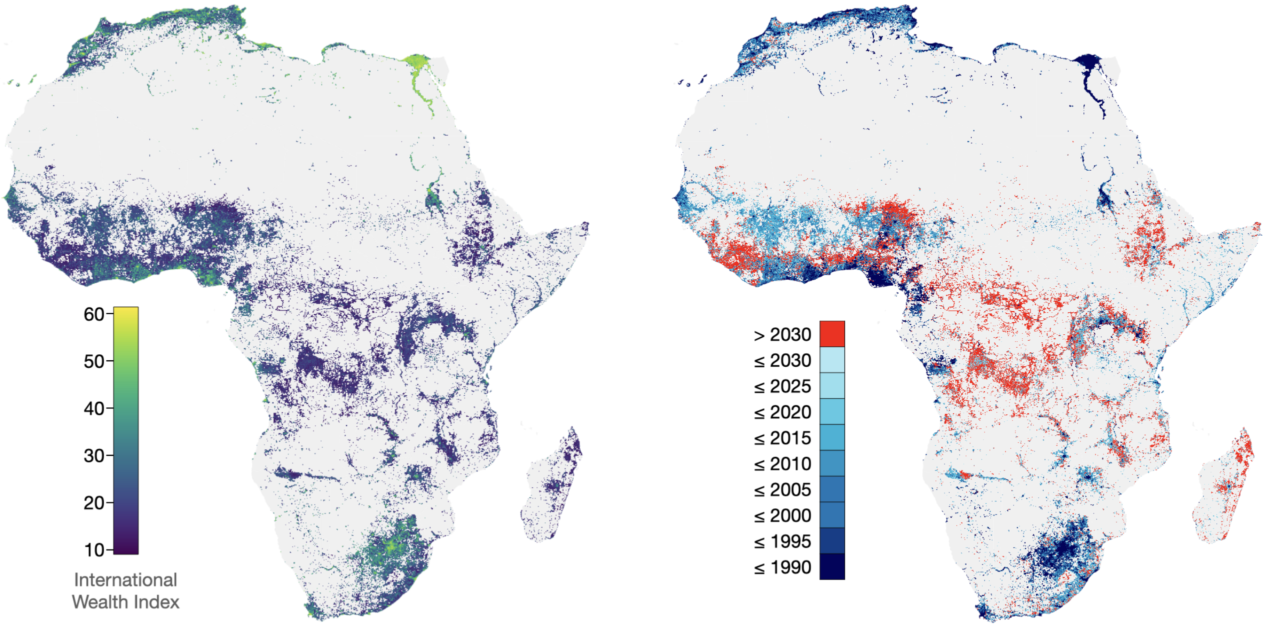
As shown in Figure 4a, even if our models do not explicitly have access to locality, they still implicitly distinguish between urban and rural neighborhoods. Although this separation accounts for a part of the explained variation, we are still able to distinguish well between poor and wealthy clusters within these two localities.

Although the temporally aware models perform better in both urban and rural neighborhoods compared to the SINGLE-FRAME model, these improvements tend to be larger in urban neighborhoods (decrease in RMSE by 2.62 compared to 0.58 for rural), as shown in Figure 4b. We can see a generally larger improvement for wealthier neighborhoods. Additionally, in rural neighborhoods, the temporally aware model notably excels in the poorest locations. For both urban and rural neighborhoods, we can see that the SINGLE-FRAME model actually performs better on neighborhoods with a wealth level close to the median (46.8 for urban and 18.5 for rural). This suggests that our SINGLE-FRAME model has a higher bias, but low variance. Nonetheless, the MULTI-FRAME models performing better in urban areas is expected, following our overarching motivation of making EO-ML methods with a temporal component. This sort of change is more visible in urban areas where roads, buildings, and infrastructure are constructed at a higher rate.

We unpack this argument—that MULTI-FRAME models perform better in urbanized areas—through a qualitative analysis of locations where the biggest and smallest improvements are made when going from SINGLE-FRAME to MULTI-FRAME. Figure 4c shows two archetypal locations where performance improved and deteriorated when switching from SINGLE-FRAME to TEN-FRAME model. The most improved prediction, the top images, is an urban neighborhood where major urbanization has occurred. The pattern is similar for other neighborhoods with major performance gains. Conversely, areas with the least improvement are typically rural with limited human activity and exhibit little to no observable change in their image sequences. Interestingly, several cases of significant performance degradation (e.g. bottom row in Figure 4c) consist of desert patches where sand dunes slowly "drift" over the years. The MULTI-FRAME models severely overestimate the wealth of these areas. With some speculation, this could be attributed to the temporally aware model picking up on large-scale changes correlating with higher wealth and confusing the movement of dunes with e.g. construction of infrastructure.

5 Applications

One use-case of our models is to create IWI maps for a variety of downstream scientific and policy tasks [Jerzak *et al.*, 2022; Jerzak *et al.*, 2023a; Jerzak *et al.*, 2023b; Burke *et al.*, 2021]. To generate this data, we used our top model to estimate IWI for all populated areas in Africa, given by population-rasters from Tatem et al. [2017]. This was done by arranging image patches of the same 6.72 x 6.72 km as was used for training. Then, we let the model predict IWI at all ten time-frames, from 1990 to 2019. The results for the last time-frame can be found in Figure 5a. As discussed in Appendix A.5, we find that these maps correlate well with national prosperity indicators such as the Human Development Index ($r^2 = 0.54$). Additionally, we make the data publicly available through the repository Harvard Dataverse.



(a) Map of predicted IWI values

(b) Map of predictions for when areas escape poverty

Figure 5: Map predictions. (a) Map showing the predicted IWI in the last of ten time-frames (2017-2019). The brighter the color, the wealthier the area. No predictions were made for the uninhabited areas in gray. (b) Map showing the time when our linear-regression model estimates an area first achieved an IWI above 20. Areas in red not achieve this level by 2030 and are therefore unlikely to make the SDG 1.

Escaping poverty One downstream task is to forecast neighborhood-level health and living conditions. The first of the UN's Sustainable Development Goals (SDGs) is to "End poverty in all its forms everywhere" by the year 2030. Nonetheless, because of previously lacking neighborhood-level data, there is a lack of forecasts that help policymakers monitor progress toward this goal. Our maps contribute to changing this situation; and using them, we fitted a linear regression model in each 6.72×6.72 km patch over time. Our fitted model predicts the future wealth development of each neighborhood, shown in Figure 5b. When setting a threshold of IWI $\geq$ 20, our forecasts reveal that from 2020 to 2030 areas that today house about 9.3% (84.8 million) of Africa's population manage to leave poverty while areas housing 15.5% (142 million) will fail to reach that threshold. Although improvements can be made to our forecasting analysis and this simplified model does not account for population growth, urban expansion or migration, it nonetheless demonstrates how our data can be used.

6 Discussion and Conclusion

State-of-the-art EO-ML models have shown that poverty levels in Africa may be imputed with acceptable precision from satellite imagery [Burke *et al.*, 2021], but they lack awareness of historical context as they rely on single-time satellite snapshots. We improve these models by infusing them with an LSTM layer and the ability to accept a multi-temporal sequence of images as input. As a result, our models gain an increased capacity to estimate health and living conditions. A critical use case of our EO-ML models is generating long-term geo-temporal data. We demonstrated this by forecasting which African neighborhoods will escape poverty, finding that 15.5% are unlikely to do so by 2030. Our findings suggest that EO-ML modeling can likely benefit other sustainable development goals as well, from monitoring climate change [Daoud *et al.*, 2016; Shiba *et al.*, 2022] to urban planning [Kino *et al.*, 2021].

Our method would benefit from addressing several key limitations in future iterations. First, future research would likely benefit from incorporating additional geo-temporal indicators (e.g., weather, roads) as well as utilizing geostatistical methods to improve estimation [González *et al.*, 2016]. Second, our EO-ML methods rely on LSTMs to capture temporal relations, yet other architectures (e.g., transformers) could further improve estimations. Third, the resolution of time frames is key. Using three-year windows to handle cloud-induced missing pixels may not be optimal for analyzing changes in IWI. In addition, a problem with using satellite images for studying poverty is that features that are likely to be picked up by a model (housing, roads, infrastructure, etc.) change slowly compared to human economic activity. An event like the coronavirus pandemic (which occurred outside our studied time period) is estimated to have increased the number of people living in extreme poverty by 119 million people [Mahler *et al.*, 2021]. But the pandemic did not inflict major change to the earth-orbital appearance of infrastructure, roads, and housing. Capturing change induced by a fast-moving and invisible-for-a-satellite event like a pandemic is an inherent limitation of any EO-ML method. To overcome this, one would need to incorporate additional sources of information, e.g. mobile-phone data [Blumenstock, 2018].

Acknowledgments

The computations and data handling were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) partially funded by the Swedish Research Council (SRC) through grant agreements no. 2022-06725 and no. 2018-05973. The AI and Global Development Lab is mainly funded by SRC grant no. 2020-03088 and 2020-00491.

We would also like to thank Fredrik Kahl, Ida Häggström, and the Computer Vision Group at Chalmers.

References

[Blumenstock, 2018] Joshua E. Blumenstock. Estimating Economic Characteristics with Phone Data. *AEA Papers and Proceedings*, 108:72–76, 2018.

[Burgert *et al.*, 2013] Clara R. Burgert, Josh Colston, Thea Roy, and Blake Zachary. Geographic displacement procedure and georeferenced data release policy for the demographic and health surveys. Technical report, ICF International, Calverton, Maryland, USA, 2013.

[Burke *et al.*, 2021] Marshall Burke, Anne Driscoll, David B. Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, March 2021.

[Chi *et al.*, 2022] Guanghua Chi, Han Fang, Sourav Chatterjee, and Joshua E. Blumenstock. Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3):e2113658119, January 2022.

[Daoud and Dubhashi, 2023] Adel Daoud and Devdatt Dubhashi. Statistical Modeling: The Three Cultures. *Harvard Data Science Review*, 5(1), jan 26 2023. https://hdsr.mitpress.mit.edu/pub/uo4hjcx6.

[Daoud *et al.*, 2016] Adel Daoud, Björn Halleröd, and Debarati Guha-Sapir. What is the association between absolute child poverty, poor governance, and natural disasters? a global comparison of some of the realities of climate change. *PLOS ONE*, 11(4):1–20, 04 2016.

[Daoud *et al.*, 2023] Adel Daoud, Felipe Jordán, Makkunda Sharma, Fredrik Johansson, Devdatt Dubhashi, Sourabh Paul, and Subhashis Banerjee. Using satellite images and deep learning to measure health and living standards in india. *Social Indicators Research*, Apr 2023.

[Daoud, 2011] Adel Daoud. *Scarcity, Abundance, and Sufficiency: Contributions to Social and Economic Theory*. Gothenburg Studies in Sociology. Gothenburg Studies in Sociology, Department of Sociology & Geson Hyltetryck, Gothenburg, 2011.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[DHS, 2022] DHS. Demographic and Health Surveys website. www.dhsprogram.com, 2022. Accessed: 2022-12-30.

[Ekbrand, 2019] Hans Ekbrand. DHSharmonisation. https://bitbucket.org/hansekbrand/dhsharmonisation/, 2019. Accessed: 2022-12-13.

[Elvidge *et al.*, 2017] Christopher D Elvidge, Kimberly Baugh, Mikhail Zhizhin, Feng Chi Hsu, and Tilottama Ghosh. Viirs night-time lights. *International Journal of Remote Sensing*, 38:5860–5879, 2017.

[Glorot and Bengio, 2010] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[González *et al.*, 2016] Jonatan A. González, Francisco J. Rodríguez-Cortés, Ottmar Cronie, and Jorge Mateu. Spatio-temporal point process statistics: A review. *Spatial Statistics*, 18:505–544, November 2016.

[Gorelick *et al.*, 2017] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 2017.

[Groves and Lyberg, 2010] Robert M. Groves and Lars Lyberg. Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, 74(5):849–879, January 2010.

[Halleröd *et al.*, 2013] Björn Halleröd, Bo Rothstein, Adel Daoud, and Shailen Nandy. Bad Governance and Poor Children: A Comparative Analysis of Government Efficiency and Severe Child Deprivation in 68 Low- and Middle-income Countries. *World Development*, 48(0):19–31, 2013.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[Henderson *et al.*, 2012] J. Vernon Henderson, Adam Storeygard, and David N. Weil. Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028, April 2012.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Hsu *et al.*, 2015] Feng-Chi Hsu, Kimberly E Baugh, Tilottama Ghosh, Mikhail Zhizhin, and Christopher D Elvidge. Dmsp-ols radiance calibrated nighttime lights time series with intercalibration. *Remote Sensing*, 7:1855–1876, 2015.

[Jean *et al.*, 2016] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353, 2016.

[Jerzak *et al.*, 2022] Connor T. Jerzak, Fredrik Johansson, and Adel Daoud. Estimating Causal Effects Under Image Confounding Bias with an Application to Poverty in Africa. *arXiv:2206.06410 [cs.LG]*, June 2022.

[Jerzak *et al.*, 2023a] Connor T. Jerzak, Fredrik Johansson, and Adel Daoud. Image-based treatment effect heterogeneity. *Forthcoming in Proceedings of the Second Conference on Causal Learning and Reasoning (CLeaR), Proceedings of Machine Learning Research (PMLR)*, 2023.

[Jerzak *et al.*, 2023b] Connor T. Jerzak, Fredrik Johansson, and Adel Daoud. Integrating earth observation data into causal inference: Challenges and opportunities. *arXiv:2206.06410 [cs.LG, stat.AP]*, January 2023.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 12 2014.

[Kino *et al.*, 2021] Shiho Kino, Yu-Tien Hsu, Koichiro Shiba, Yung-Shin Chien, Carol Mita, Ichiro Kawachi, and Adel Daoud. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM - Population Health*, 15:100836, September 2021.

[Lavrakas, 2008] Paul Lavrakas. *Encyclopedia of Survey Research Methods*. Sage Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America, 2008.

[Li *et al.*, 2019] Ang Li, Ola Spyra, Sagi Perel, Valentin Dalibard, Max Jaderberg, Chenjie Gu, David Budden, Tim Harley, and Pramod Gupta. A generalized framework for population based training. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 1791–1799, New York, NY, USA, 2019. Association for Computing Machinery.

[Mahler *et al.*, 2021] Daniel Gerszon Mahler, Nishant Yonzan, Christoph Lakner, R. Andres Castaneda Aguilar, and Haoyu Wu. Updated estimates of the impact of covid-19 on global poverty: Turning the corner on the pandemic in 2021? *World Bank Data Blog*, 2021.

[Rolf *et al.*, 2021] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1):4392, July 2021.

[Shiba *et al.*, 2022] Koichiro Shiba, Adel Daoud, Hiroyuki Hikichi, Aki Yazawa, Jun Aida, Katsunori Kondo, and Ichiro Kawachi. Uncovering Heterogeneous Associations Between Disaster-Related Trauma and Subsequent Functional Limitations: A Machine-Learning Approach. *American Journal of Epidemiology*, page kwac187, October 2022.

[Smits and Steendijk, 2015] Jeroen Smits and Roel Steendijk. The international wealth index (iwi). *Social Indicators Research*, 122:65–85, 2015.

[Tatem, 2017] Andrew J. Tatem. Worldpop, open data for spatial demography. *Scientific Data*, 4(1):170004, Jan 2017.

[World Bank, 2022] World Bank. Poverty and inequality platform (version 20220909_2011_02_02_prod). www.pip.worldbank.org, 2022. Accessed on 2022-12-13.

[Yeh *et al.*, 2020] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 11, 2020.