Understanding Economic Development in Rural Africa Using Satellite Imagery, Building Footprints and Deep Models

Amna Elmustafa, Erik Rozi, Yutong He, Gengchen Mai, Stefano Ermon, Marshall Burke, David Lobell
African Institute for Mathematical Sciences, Senegal
aelmustafa@aimsammi.org
Department of Computer Science, Stanford University, Stanford, CA, USA
{erikrozi,kellyyhe,maigch,ermon}@cs.stanford.edu
Department of Earth System Science, Stanford University, Stanford, CA,USA
{mburke,dlobell}@stanford.edu

ABSTRACT

Recent advancements in machine learning enable cost effective methods for understanding societal and economic activities in developing countries using publicly available satellite imagery. However, this progress remains stagnant in rural areas where the largest population under poverty line resides. In this work, we explore deep models' performance in rural areas in Africa and investigate methods that improve the performance. We argue that the geographic displacement noise present in ground surveys for anonymization purposes causes misalignments between input imagery and labels and therefore hampers accuracy, which exacerbates in rural areas. We then propose to incorporate building footprints data and a novel self-attention mechanism to provide more robust and accurate predictions of socioeconomic development. We test our framework against three socioeconomic measures in 21 African countries. Our best models outperform previous baselines in most of these tasks.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Applied computing** → *Economics.*

KEYWORDS

Deep Learning, Convolutional Neural Network, Supervised regression, Prediction in Rural Area, Poverty, Development, Satellite Imagery

ACM Reference Format:
Amna Elmustafa, Erik Rozi, Yutong He, Gengchen Mai, Stefano Ermon, Marshall Burke, David Lobell. 2022. Understanding Economic Development in Rural Africa Using Satellite Imagery, Building Footprints and Deep Models. In *The 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22), November 1–4, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3557915.3561025

1 INTRODUCTION

The standard ways of collecting socioeconomic data to track sustainability measures depend on doing ground surveys or censuses. These methods require considerable technical expertise as well as financial resources, hindering developing countries' ability to track key socio-economic indicators of sustainable development [1].

On the other hand, alternative sources of data such as satellite images [9, 16], mobile phones and social media data [1], and Wikipedia articles [12] are freely available in abundance. With the recent advancements in machine learning, leveraging these publicly available sources of data for developmental measures becomes possible. These methods better generalize through time and space, inform in the end the policy-making processes in low resource settings and enable a more sustainable way to track social-economic activities in developing countries.

While most of previous works [7, 9, 11, 12] evaluated their models on *average* performance (e.g., accuracy or mean-squared error), 60% of the population in the developing countries resides in rural areas where the poverty rate is significantly higher [15], which indicates the importance of separate evaluation of the performance on rural sub-populations. Some prior work [10, 16] tested their frameworks on rural sub-populations, and an almost 40% decrease on Pearson's r^2 was reported compared to the performance on overall population (see Figure 1). Therefore, in this work, we focus on evaluating and improving the performance of satellite-based predictions on socio-economic indicators over rural sub-populations.

To trace the root of the performance degradation, Koh et al. [10] formulated it as a domain shift problem between the training and testing set. However, in Figure 1 we show that even if we focus both training and testing only on the rural sub-population, the performance gap between urban and rural regions is still prominent, which explains the insignificance of the performance improvement provided by the training schemes in [10].

We hypothesize that the performance drop of deep models in rural areas can be caused by the geographic displacement noise used for anonymization purposes when collecting ground surveys, which results in misaligned datasets: the geo-locations of the satellite images correspond to the true geo-locations of the survey data plus some random noise. Hence, the mapping between input images and output labels is incorrect. Because residents in rural areas usually live further apart, survey makers often provide larger geographic displacement noise in rural areas than in urban areas [2]. In demographic and health surveys (DHS), urban areas shift is up to 2 km, while the maximum jitter is 10 km for rural data (see example in Figure 1). This larger noise presents greater challenge for prediction in rural regions.

SIGSPATIAL '22, November 1–4, 2022, Seattle, WA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9529-8/22/11.
https://doi.org/10.1145/3557915.3561025

[1] https://www.premiumtimesng.com/news/headlines

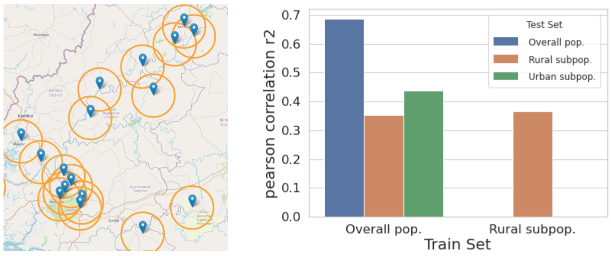


Figure 1: left-image: Geographic displacement noise in selected locations in Kambia, Sierra leon. Circles represent the radius of the noise; right-image: Test performance of previously used deep models in urban-rural sub-population [16]. *x-axis* represents which population/sub-population we train on, and *y-axis* represents test performance(r^2) among urban, rural and overall test sets, each specified by a different color.

Motivated by the correlations between building objects and different sustainability measures [11] as well as the recently publicly released labelled building footprints across Africa [13], we propose to use building footprints as a separate input signal that can provide additional (potentially, more accurate) information about the socioeconomic and geographical characteristics of a population. Figure 2 shows an example of building density differences. Compared with a model that uses satellite imagery alone, we expect that adding building footprints can increase both the overall performance and the rural areas’ performance by decreasing the effect of displacement noise.

We also propose a model architecture inspired by the recent advancements in attention mechanism [4]. Our model first divides the input image into patches and then extracts features from each patch. An attention layer then takes in the visual features of all patches and learns to better focus/attend on the informative portion of the images. By incorporating attention mechanism, we expect the model to identify the target settlements even in the presence of large displacement noise.

We evaluate our models on various socioeconomic indicators across 21 African countries. Our best models outperform previous baselines especially in rural areas, with an improvement in Pearson correlation metric that varies from 3-15%, on wealth index, sanitation index and women education attainment index.

2 METHOD

We propose to incorporate two novel approaches for more robust predictions in rural areas: (1) Introducing **building footprints** as a new source of data for the predictions; (2) using **self-attention mechanism** over image patches while including a larger neighbourhood from the image center as input.

2.1 Building footprints as a signal for predicting sustainability measures

A "rural" area according to DHS surveys is defined based on measurements of population density, the infrastructure of the area, and occupation of inhabitants. Building footprints, which are polygon indicators that represent the location and the shape of buildings on satellite images, provide clearer information about human activities and population that relates to both socioeconomic metrics and rural/urban indications than satellite images alone. Moreover, we observe that since it is possible to shift the central location but still stay in an empty area without detecting new buildings, including building footprints in rural areas predictions can help the jitters to be less prominent (example in Figure 2). Because the features detected in the shifted images are unaffected, the negative effect of the misalignment becomes less significant. Hence, incorporating building footprints in these tasks improves the displacement noise robustness, and in turn, can help the prediction for both the whole population and the rural sub-population.

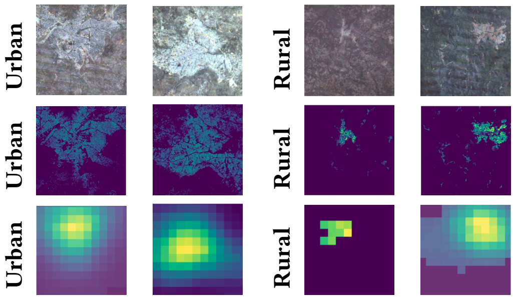


Figure 2: Rural vs Urban image pairs. (Top: Satellite, Middle: Building, Bottom: Nightlight)

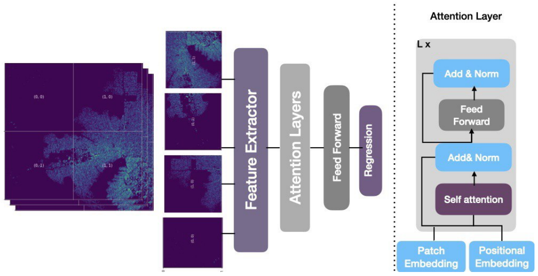


Figure 3: Model architecture with attention layers. (Left: Full model architecture, Right: Details about the attention layer)

Recently, Sirko et al. [13] have released a publicly available dataset of around 516-Million Africa-wide buildings bounding box polygons. To adapt these footprints to our pipeline, we first map the surveys locations to the polygon locations, and then we rasterize the building footprints into images with 30m spatial resolution to align with the other satellite bands we use, where pixel values represent building area ratio. The resulting images have one additional building footprint channel. Figure 2 shows examples of this newly added one-channel images of building, paired with their corresponding RGB and nightlight bands, in rural/urban areas.

2.2 Attention-based Model

2.2.1 Encoding the image patches. As shown in Figure 3, we develop an image patch feature extractor that encodes image patches into embeddings before inputting them into the attention layers. Given an image $X \in \mathbb{R}^{C\times H\times W}$, where H, W, and C indicate image height, width, and number of channels, we first divide it into N patches with size $P\times P$ where $P \leq H, W$. Here, $N = ((H-P)/S+1)\times((W-P)/S+1)$ where S indicates stride. Notice that patches can partially overlap with each other. Each image patch $X_p \in \mathbb{R}^{C\times P\times P}$ is then further encoded into a D dimensional image patch embedding (see Equation (1)), using a pretrained ResNet18. We use a 2D position encoder to represent the position of each patch into a D dimensional embedding and add to each patch embedding (i.e., element-wise summation). Next, these N image patch embeddings are further fed into self-attention layers to aggregate them into one single image embedding.

2.2.2 Attention layers. We incorporate self-attention to better locate the important signals in the inputs. Following [4, 14], we use L number of alternating self-attention (SA) (Equation 1, 2, and 3), as well as feed forward layers (MLP) comprised of two linear layers with GELU non-linearity, layer normalization (LN), and residual connection before and after each layer (See Equation 4, 5, 6, 7). The resulting N feature maps are aggregated to a global average (see Equation 8). This global feature map is then fed to the final fully connected layer to predict the specific sustainability index.

$$[q, k, v] = ZP_{qkv}, \quad P_{qkv} \in \mathbb{R}^{D\times D}, \quad Z \in \mathbb{R}^{N\times D}. \tag{1}$$

$$A = \text{softmax}\left(\frac{qk^T}{\sqrt{D}}\right), \qquad A \in \mathbb{R}^{N\times N}. \tag{2}$$

$$\mathrm{SA}\,(Z) = Av. \tag{3}$$

$$Z^0 = \left[E\left(x_p^1\right); E\left(x_p^2\right); \cdots; E\left(x_p^N\right)\right] + E_{pos}, \qquad E \in \mathbb{R}^D, \quad E_{pos} \in \mathbb{R}^D. \tag{4}$$

$$Z_\ell^1 = \mathrm{SA}\left(\mathrm{LN}\left(Z_{\ell-1}^2\right)\right) + Z_{\ell-1}^2; \qquad \ell \in 1, \cdots, L; \quad Z_\ell^1 \in \mathbb{R}^{N\times D}. \tag{5}$$

$$Z_\ell^2 = \mathrm{MLP}\left(\mathrm{LN}\left(Z_{\ell-1}^1\right)\right) + Z_{\ell-1}^1; \qquad \ell \in 1, \cdots, L; \quad Z_\ell^2 \in \mathbb{R}^{N\times D}. \tag{6}$$

$$y = \mathrm{LN}\left(Z_L^2\right). \tag{7}$$

$$y_{agg} = \frac{1}{N}\sum_{i=1}^{N} y_i, \qquad y_{agg} \in \mathbb{R}^D. \tag{8}$$

3 EXPERIMENTS

3.1 Datasets

3.1.1 Demographic and Health Surveys (DHS). We derive our labels/ground truth data from the DHS surveys [3] curated by [16, 17]. DHS surveys are nationally-representative ground surveys with different questionnaires that track demographic and socioeconomic indicators. Often times, they also include geographic information (latitude and longitude) about each surveyed cluster. These clusters usually represent randomly selected households from each enumeration area in census files, and are labeled either rural or urban. We focus on questionnaires related to wealth, health and education indices collected from surveys conducted between 2009 and 2016 in 21 African countries, summing all up to 18503 clusters.

3.1.2 Satellite imagery and building footprints . We use the publicly available google earth engine tool to collect the satellite imagery and the building data. This tool can be used to combines multiple satellite images with other geo-spatial data [6]. By using Google Earth Engine, we geographically align our 18503 DHS clusters/locations with the corresponding satellite images and building footprints. We have 3 types of input bands: Multi-spectral, Nightlight and Building footprints.

Multispectral (MS) bands. MS bands are 7 bands collected following the same procedure in [16], from a 3-year cloud-free median composite of surface reflectance values in Landsat 5,8 and 7. All the bands have 30 m per pixel resolution, representing RED, GREEN, BLUE, NIR (Near Infrared), SWIR1 (Shortwave Infrared 1), SWIR2 (Shortwave Infrared 2), and TEMP1 (Thermal).

Nightlights (NL) band. Also following [16], NL bands are composed of 3 years composite of night-time lights recorded from 2 sets of satellite DMSP and VIIRS [5, 8]. Each of them captures data from different years range (2009-2011 and 2012-2016 respectively).

Building footprints (building). Using mean reduction, we reduce pixel count in [13] to 30m per pixel, after filtering bounding boxes with less confidence score (<0.7).

In the end, our satellite imagery can have 7 MS bands, 1 NL band and 1 building band. Each input can be used separately or in combination with other bands. We refer to each input combination as (NL+MS), (NL+MS+building), (NL+building), and (building+MS).

3.2 Baselines and Evaluation metrics

In previous work, the best model we compare against for the wealth index task is the ResNet18[16] with MS+NL input, while KNN with center NL pixel input was the best for other tasks. For evaluation, we use the squared Pearson correlation coefficient both for whole and sub-populations(r^2), which is known to capture how variance in the ground truth label is explained by model predictions.

3.3 Experiments on wealth, health and education indices prediction

Experiments are carried out to investigate how different inputs or whether using attention layers over larger image affects model performance in rural settings. Results for wealth index, sanitation index, and women education index are shown in Table 1, 2, and 3 respectively. Moreover, to investigate more how attention weights are improving performance, we compare its performance against taking the global average of the patches as shown in table 1

3.4 Summary on DHS indices prediction

Generally speaking, adding building data as model input is a key to improving model performance in rural areas with 5-13% increase in r^2, compared to using NL or MS bands. Correlating education and sanitation indices with satellite images is more difficult compared to wealth index and they are more prone to noise. We hypothesize that this is due to the fact that they are calculated from a single variable in DHS surveys, unlike wealth index which is aggregated from multiple variables. Better performance may be achieved by supplementing building data with other sources of input such as street imagery [11] or Wikipedia articles [12], but we leave this as future work.

Table 1: Results on wealth index prediction task. "GA" indicates global average pooling; "SA" denotes self-attention layers.

Resnet18 baseline			
input band	r^2	r^2 **rural**	r^2 **urban**
MS+NL	0.69	0.35	0.33
Ablation on input type - ResNet18			
input band	r^2	r^2 **rural**	r^2 **urban**
MS	0.52	0.15	-0.05
NL	0.67	0.33	0.3
Building	0.67	0.33	0.3
MS+NL ([16])	0.69	0.35	0.33
MS+Building	0.69	0.35	0.33
NL+Building	0.71	0.4	0.47
MS+NL+Building	**0.72**	**0.41**	**0.48**
Ablation on model- NL+building input			
model	r^2	r^2 **rural**	r^2 **urban**
ResNet+SA	**0.72**	**0.43**	**0.46**
ResNet+GA	0.68	0.39	0.43

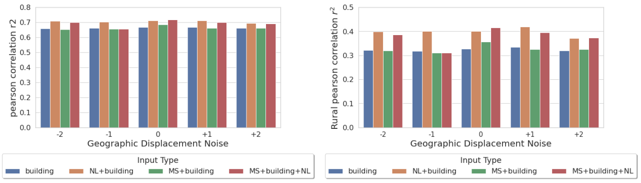
Table 2: Results on the sanitation index prediction task.

KNN Baseline			
input band	r^2	r^2 **rural**	r^2 **urban**
NL center pixel	0.39	0.069	0.22
ablation on input type-ResNet18			
input band	r^2	r^2 **urban**	r^2 **urban**
NL	0.24	0.01	0.11
Building	**0.41**	**0.13**	0.17
NL+building	0.39	0.104	0.16
NL+building+MS	0.29	0.03	0.17
Model-NL+building			
model	r^2	r^2 **rural**	r^2 **urban**
resnet+SA	**0.43**	**0.134**	0.22

Table 3: Results on the women education attainment index prediction task.

KNN Baseline			
input band	r^2	r^2 **rural**	r^2**urban**
NL center pixel	0.14	0.002	0.015
Ablation on input type-resnet18 model			
input band	r^2	r^2 **rural**	r^2 **urban**
NL	0.15	0.01	0.019
Building	0.24	0.06	0.1
NL+building	0.25	0.07	0.1
NL+building+MS	**0.25**	**0.1**	**0.15**
Model-NL+building input			
model	r^2	r^2 rural	r^2 urban
resnet+SA	0.23	0.07	0.08

Adding the image patch self-attention layer provides 4% r^2 improvement on the sanitation index prediction task and 3% in the wealth index task, but only marginal advantages on women education index prediction task. We expect further improvement when using higher resolution imagery in future work, as self-attention benefits from higher quality inputs when extracting more fine-grained information.



(a) Overall perform. **(b) Rural areas perform.**

Figure 4: The model performance against different levels of geographic displacement noises with building input (a,b). "+" and "-" indicates geographic displacement to the left or right.

3.5 Experiments on robustness to different levels of Noise

The goal of this section is to investigate how robust models with different bands are to different levels of geographic displacement noise. We perform detailed comparison in Figure 4.

We simulate the noise by shifting the image's center within its neighbourhood of 2 km, by a distance of either 1 or 2 Km, to the left (+) or right (-) direction. We use wealth index labels for these experiments and compare robustness before and after adding the building footprint data.

We can see that noise intensely affects the MS bands by causing an almost 10% decrease for different noise levels. As expected, this effect can be compensated for by combining the additional information from the building footprint band.

4 CONCLUSION

We propose two novel ways to mitigate the negative effect of the geographic displacement noise from ground surveys on deep remote sensing based model in rural Africa: (1) We use building footprints as an additional signal to provide more robust input to the deep models and (2) we incorporate self-attention mechanism to better extract visual features. We assess our performance on three different tasks in 21 African countries. On most of these tasks, we achieve significant improvements. We hope that our work can attract more attention toward the overlooked rural areas progress toward sustainable development goals, and encourage more research on debiasing the deep models performance on underrepresented areas, which in a way can lead to a fairer distribution of the limited resources in the developing world.

REFERENCES

[1] Joshua Blumenstock et al. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350, 6264 (Nov. 2015), 1073–1076.

[2] Clara R Burgert et al. 2013. *Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys.* ICF International.

[3] DHS. 2019. DHS Demographic and Health Surveys 1996-2019. Funded by USAID. https://dhsprogram.com/.

[4] Alexey Dosovitskiy et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR 2021*.

[5] Christopher D Elvidge et al. 2017. VIIRS night-time lights. *International Journal of Remote Sensing* 38, 21 (2017), 5860–5879.

[6] Noel Gorelick et al. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment* 202 (2017), 18–27.

[7] Andrew Head, Mélanie Manguin, Nhat Tran, and Joshua E Blumenstock. 2017. Can human development be measured with satellite imagery?. In *Ictd*. 8–1.

[8] Feng-Chi Hsu, Kimberly E Baugh, Tilottama Ghosh, Mikhail Zhizhin, and Christopher D Elvidge. 2015. DMSP-OLS radiance calibrated nighttime lights time series with intercalibration. *Remote Sensing* 7, 2 (2015), 1855–1876.

[9] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (Aug. 2016), 790–794.

[10] Pang Wei Koh et al. 2021. *WILDS: A Benchmark of in-the-Wild Distribution Shifts.* Technical Report arXiv:2012.07421. arXiv. http://arxiv.org/abs/2012.07421

[11] Jihyeon Lee et al. 2021. Predicting Livelihood Indicators from Community-Generated Street-Level Imagery. In *AAAI 2021*, Vol. 35. 268–276.

[12] Evan Sheehan et al. 2019. Predicting Economic Development using Geolocated Wikipedia Articles. In *ACM SIGKDD 2019*. 2698–2706.

[13] Wojciech Sirko et al. 2021. Continental-Scale Building Detection from High Resolution Satellite Imagery. *arXiv preprint arXiv:2107.12283* (2021).

[14] Ashish Vaswani et al. 2017. Attention is all you need. *NeurIPS 2017* 30 (2017).

[15] world bank. 2020. World Bank Rural population - Sub-Saharan Africa. https://data.worldbank.org/.

[16] Christopher Yeh et al. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications* 11, 1 (Dec. 2020), 2583.

[17] Christopher Yeh et al. 2021. SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning. In *NeurIPS 2021*.