Series Name: WPS
WPS Paper No. 219
Issue Date: 1/27/23

High-resolution rural poverty mapping in Pakistan with ensemble deep learning

Felix S. K. Agyemang; Rashid Memon; Levi John Wolf, and Sean Fox

High-resolution rural poverty mapping in Pakistan with ensemble deep learning

Felix S. K. Agyemang [a]; Rashid Memon [b]; Levi John Wolf [c]; Sean Fox [c]

[a] Department of Planning and Environmental Management, University of Manchester, UK

[b] Department of Economics, Lahore University of Management Sciences, Pakistan

[c] School of Geographical Science, University of Bristol, UK

Abstract

High resolution poverty mapping supports evidence-based policy and research, yet about half of countries lack the requisite survey data to generate useful poverty maps. To overcome this challenge, new non-traditional data sources and deep learning techniques are increasingly used to create small-area estimates of poverty in low- and middle-income countries (LMICs). Convolutional Neural Networks (CNN) trained on satellite imagery are one of the most popular and effective approaches in this literature. However, the spatial resolution of poverty estimates has remained quite coarse, particularly in rural areas which are critical for governments to support. To resolve this, we use an ensemble transfer learning approach involving three CNN models to predict chronic poverty at a finer 1 km^2 scale in rural Sindh, Pakistan. We train the model with spatially noisy georeferenced household survey containing poverty scores for 1.9 million anonymized households in Sindh Province using publicly available inputs, including daytime and nighttime satellite imagery and accessibility data. Results from rigorous cross-validation and ground truthing of predictions with an original survey suggest the model performs well in identifying the chronic poor in both arid and non-arid regions, outperforming previous studies in key accuracy metrics. Our inexpensive and scalable approach could be used to improve poverty targeting in low- and middle-income countries.

Keywords: Deep Learning; Convolutional Neural Network; Poverty Mapping; Low and Middle Income Countries; Pakistan.

Introduction

The impact of Covid-19 on lives and livelihoods has accelerated social protection support efforts by governments and non-governmental organisations across the globe. High resolution poverty mapping supports evidence-based policy and research (Yeh et al 2020), yet an alarming half of all countries do not have access to sufficient data produce such maps (Serajuddin et al., 2015). Census data has been the traditional source of generating economic data about populations in many developing countries. Yet, in addition to being expensive, census in most countries are not conducted frequently (Lucci et al., 2018; Onda et al., 2019), hampering its utility in rapidly evolving developing countries. The Demographic and Health Surveys (DHS) has emerged as a popular source for generating nationally and sub-nationally representative socio-economic and health data, however it has a small sample, and its spatial coverage is rather sparse in many countries. Significantly expanding the coverage of DHS data will be costly and challenging for most low and middle income countries (Jerven, 2014).

To address these challenges researchers have explored alternative and less costly approaches to estimating economic activity at subnational levels. Early efforts used luminosity from nighttime lights (NTL) as a proxy for measuring economic activity (Henderson et al., 2012; Bleakley and Lin, 2012; Engstrom et al., 2017; Watmough et al., 2019). NTL luminosity has been found to directly correlate with wage income (Mellander et al., 2015) and asset wealth (Noor et al., 2008) at various spatial scales. However, nighttime lights are highly limited in observing variations in economic activity and living standards in LMIC (Chen et al., 2011; Mellander et al., 2015; Jean et al., 2016), the areas in most need. Moreover, using commune-level dataset from Vietnam, Goldblatt et al. (2020) found daytime satellite imagery better predicts economic activities than nighttime lights.

The use of satellite imagery to map poverty and economic activity has grown in the past decade alongside improvements in machine learning and computer vision. Many of these approaches make use of deep learning techniques such as Convolutional Neural Networks (CNN). For example, Yeh at al. (2020) used CNN to map wealth across 20,000 villages in Africa; Chi et al (2022) used CNN with boosted regression trees to predict asset wealth in LMICs; Jean et al. (2016) combined CNN with ridge regression to measure consumption expenditure in five African countries; Sung et al. (2020) used the approach to estimate GDP in US counties; and Xie et al (2016) mapped poverty in Uganda with the combination of CNN and logistic regression. In addition to generating good results, most of these CNN models have been deployed in data-challenged regions, mainly in LMICs (see Chi et al., 2022; Head et al., 2017; Babenko et al., 2017; Persello and Stein, 2017; Wang et al., 2019). In addition, most models are trained on publicly available satellite imageries as inputs, including Landsat (Perez et al., 2017), Google Static Maps (Jean et al., 2016), DMSP and VIIRS (Chi et al., 2022), making them not only scalable but also inexpensive to apply in the real world.

Many of these CNN models were used to generate estimates in both urban and rural areas. Yet there are strong *a priori* reasons to expect substantial differences in the types of visual information required to accurately predict variation in economic activity or living standards between and within rural and urban areas. For example, technologies of production vary widely between agricultural and non-agricultural contexts, as do indicators of consumption (such as dwelling size). While broad variations in economic activity or welfare between urban and rural areas may be visible from the sky, it is more difficult to observe differences *within* either context from the above. Urban areas tend to be more socio-economically heterogeneous, with building features as well as morphology generally reflecting socio-economic characteristics of households (Kuffer and Barros, 2011; Tapiador et al., 2011; Wurm and Taubenböck, 2018).

Rural exhibit less architectural and morphological variation, although may have greater variation in landscapes that contain information on household living conditions. It is therefore unlikely that a single model applied to satellite imagery could reflect intra-urban *and* intra-urban variation in household welfare at a high spatial resolution. Put differently, it is not terribly difficult to make broadly accurate spatial estimates of relative living standards between urban and rural areas with satellite imagery given their distinct economic characteristics. By contrast, it is challenging to generate accurate spatial estimates at a high resolution *within* each of these contexts.

Yet there has been some progress in this area, notably the mapping of asset wealth across 20,000 villages in Africa (Yeh et al., 2020). However, the spatial resolution of Yeh et al's work, as with other studies using CNN, is coarse—especially in rural contexts (see Head et al., 2017; Xie et al., 2016; Sung et al., 2020). The resolution of Yeh et al. (2020) CNN model is 6.72km * 6.72km, which in the context of most developing countries will reflect an estimate for many rural settlements. Even though the 'micro-estimates' of wealth (Chi et al., 2022) is presented at 2.4km resolution, the underlying DHS data, the target layer or label for the CNN model, was aggregated to 4.8km grid cells in urban areas and 9.6km grid cells in rural areas. The coarseness of the existing deep learning models is largely influenced by the sources of the economic data used for training the networks. Data from the DHS, the dominant source, is spatially distorted up to 2km in urban areas and 5km in rural areas to preserve the anonymity of households. Similarly, the geographical coordinates of the Living Standard Measurement Study (LSMS) contain up to 5km noise. Existing models therefore resort to coarser spatial resolutions to reduce their sensitivity to this small-scale locational noise.

Policy makers in LMICs seeking to target livelihood interventions in rural areas at a much finer scale will have major challenges relying on existing measurements of economic well-being. Pakistan is one of such cases. Pakistan's Sindh Province, home to an estimated 48 million people, has established a Strategic Social Protection Unit (SPSU) and assigned it resources to develop a targeting strategy. The SPSU has also been tasked with identifying eligible households in rural Sindh for cash relief in response to shocks, such as Covid-19 and monsoon floods.

Building on existing efforts, we develop and train an ensemble CNN model to generate small-area estimates (1km^2) of poverty in rural Sindh to support such targeting. We utilize an extensive georeferenced household survey containing data on assets and poverty scores for 1.9 million anonymized households in Sindh province. Asset based poverty and wealth indices are generally seen as less noisy and more stable, especially in the long term, than those based on consumption (Sahn and Stifel, 2003; Filmer and Scott, 2012). Whilst the survey is comprehensive, it has significant spatial distortions making our task comparable to past studies that used noisy datasets like the DHS. Making predictions at finer resolutions such as 1km^2 is challenging, but results show the model is promising: it compares well with past studies and has decent performance in a random benchmark test.

Defining and measuring rural poverty in Pakistan

While definition and measurement of poverty remains contested (Fletchner, 2021), we use a poverty measure based on household assets, which is both conceptually robust and practical. This approach builds on the 'basic needs' concept that constitutes the primary framework for defining national poverty lines (Atkinson, 2019). Historically, the use of a basic needs approach measure poverty was limited to rich countries. From the 1980s, the institutionalization of the Living Standards Measurement Surveys (LSMS), promoted by the World Bank, regularly provided the data necessary for poverty line measurement in developing countries as well (Deaton 2003). However, these surveys are financially and technically demanding and provide data at very course resolution. As Covid-19 recently demonstrated, many policy makers in developing countries require high resolution measures of poverty that can be collected quickly, accurately and cost effectively.

"Quick and dirty" measures of poverty (Chambers, 1981), have therefore developed alongside the "long and clean" measures based on large household surveys. Participatory Poverty Assessments, for example, became very popular among NGOs after the 1970s. Based on the principles of 'optimal ignorance' (importance of knowing what is not worth knowing) and proportionate accuracy (much survey data has a degree of accuracy that is unnecessary), these measures provided a shortcut, avoiding more expensive direct and time-consuming investigations (Chambers, 1979). Since then, participatory poverty assessments have been conducted in many countries in East and South Asia, Africa and Latin America (See references in Aczona (2009), Eden et al (2019) and Gow (2019).

This kind of data can then be used to improve targeting by providing information on relative need through a 'proxy means test' (PMT) to predict whether a household is poor (i.e. in need of government support) or not. This approach is particularly valuable in LMICs with limited household data (Grosh and Baker, 1995). The World Bank, in particular, uses detailed household surveys (e.g. LSMS) to establish PMT models for individual countries (e.g. Sebastian et al 2018), which can then be used to produce household-level estimates of poverty with 'quick and dirty' data collected at higher frequency and lower cost.

One of the most popular rapid data collection methodologies is the Simple Poverty Score (SPS) developed by Schreiner (2006) with support from the Ford Foundation and Grameen Foundation, which has now been used in 63 countries (Skoufias et al 2020). The SPS is similar in approach to the USAID's Poverty Assessment Tool in method but claims a greater degree of transparency and ease of use (Schreiner, 2014). It uses a 10-question survey and weights estimated from nationally representative surveys using logistic regression.

The SPS requires information in three main areas: the location of a household, household member characteristics, and household assets such as air conditioners, refrigerators, vehicles, agricultural land and livestock. Data on these characteristics is then compressed into twelve indicators. For example, ownership of refrigerators, freezers and washing machines is lumped into one binary indicator, which takes a value of 1 if a household possesses any of the three assets. Similarly, air conditioners and heaters are compressed into one indicator. Each indicator is then assigned a weight. The exact choice of the twelve indicators and the weights assigned to them depend on the context. Logistic regression models using data from Living Standards Measurement Surveys are typically used to identify which 12 indicators are the best predictors of poverty in any given country and period (see Schreiner 2006 for details). The coefficients from the regression are then transformed into weights for each indicator. The total score, the sum of individual scores, can then be related to the probability that a household is poor by using a simple statistical table. A local pro-poor organization can then implement a small household survey based on just these 12 indicators, calculate a poverty score, and determine eligibility for a household to receive subsidized goods and services.

The SPS has received much attention from World Bank Programming as the Bank's twin goals of eliminating extreme poverty and inequality require measuring poverty rates in specific populations targeted by development programs worldwide. Considering the time and cost required for using poverty measures based on large scale survey data, and the data and technical demands of using small area estimations, the SPS has become the post popular solution for project specific poverty estimation (Skoufias et al 2020).

The development of the SPS for Pakistan using PSLM 2005/06 is documented in Schreiner (2010), and an update using the PSLM 2007/08 is documented in Hou (2009). Each household receives a score between 0 and 100, with higher scores indicating lower levels of deprivation. In January 2009, The Government of Pakistan adopted the poverty SPS as the targeting tool for the Benazir Income Support Program, the flagship cash transfer program. The cut-off score for poverty was decided as 17.5, at which 16.3 percent of families – about 5.9 million – would be covered. This cutoff was chosen to align with an estimated national poverty headcount of 17% at the time (Hou, 2009). In general, the rural poverty rate in Pakistan has always exceeded the national poverty rate by about 5 to 6 percentage points (Government of Pakistan, 2016).

The SPS on which this paper is based was collected from 1.9 million households in 14 districts in Pakistan's Sindh province, as part of the Sindh Union Council Economic Strengthening Support (SUCESS) Programme. The SUCCESS Programme covered eight out of the province's 24 districts; data were collected in an additional six were by the Government of Sindh (GoS) and Sind Rural Support Organization (SRSO) under the People's Poverty Reduction Program (PPRP). Table 1 lists the enumerated districts and presents the household count from the Population Census 2017[1], which was conducted soon after the poverty scoring.

Table 1: Simple Poverty Score Coverage in Districts

Districts	Rural HH (Census 2017)	SPS Enumeration Coverage
Badin	282909	75.1
Jamshoro	104518	74.8
Dadu	216911	78.5
Matiari	109997	70.9
Sujawal	136805	73.0
Tando AY	113185	73.0
Tando MK	104297	69.9
Thatta	150588	88.7
Kamber	155566	94.4
Larkana	142358	85.0
Mirpur Khas	205234	75.6
Umerkot	164990	77.6
Sanghar	267383	77.3
Khairpur	279258	95.7

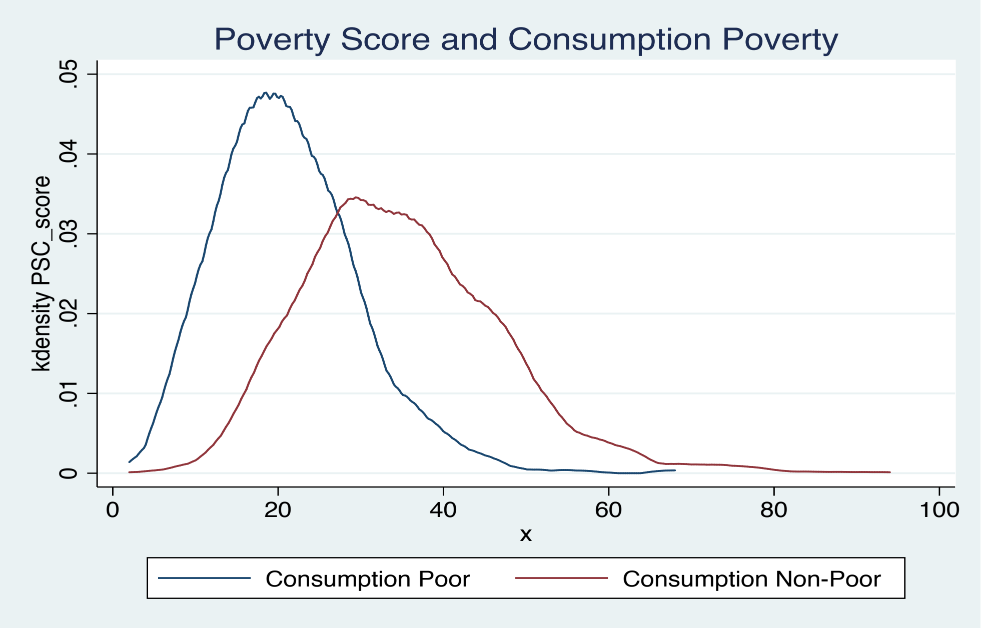
Source: SUCCESS/PPRP and authors' calculations using Population Census 2017

As can be seen in Table 1, the SPS survey coverage is very high but not complete. The data were collected through computer-assisted personal interviews (CAPI) at the doorstep of each household using Android Tablets. Household information was collected from a household member older than 18 years, with preference for the head or the spouse of the head of the household. GPS readings were taken at the end of each interview.

[1] The exact number of rural households was not available. We were calculated these using the total population, the proportion of population that was rural and the average household size in the district. (See https://www.pbs.gov.pk/sites/default/files//population_census/District%20wise%20Sindh%20TABLE%201%202017%20FINAL.pdf)

We can evaluate the relationship between household SPS scores in our sample with the consumption-based measure used for Pakistan's official poverty line (GOP, 2016b) using data from the Household Integrated Economic Survey (HIES) from 2015-16. Figure 1 shows the distribution of the poverty score for the consumption poor and consumption non-poor using the government definition. There is overlap between poverty scores of the poor and non-poor, which means there is a risk of inclusion errors at low poverty scores (i.e., counting the non-poor as poor due to a low SPS). Above a SPS of approximately 70 the probability of being poor drop to zero; the probability of being non-poor is 'reasonably' high above a poverty score of approximately 30.

Fig 1: The Simple Poverty Score and consumption poverty



Source: Authors' estimations using HIES 2015-16

Given the different dimensions of life measured in the SPS, it is natural to explore relationship between the SPS and a measure of multidimensional poverty, which has gained prominence in subnational poverty research and policy in recent years. The SPS explicitly uses poverty on non-monetary dimensions aims to predict consumption / basic-needs poverty. Multidimensional poverty, on the other hand explicitly recognizes that poverty along social dimensions *may not necessarily* accompany reductions in consumption poverty. Vision 2025, which institutionalized a multidimensional poverty index (MPI) to informs policymaking in Pakistan, was designed with the explicit aim of balancing progress on monetary measures of poverty with that in the social dimension (Government of Pakistan, 2016a). For example, a household is considered deprived in education if a child is not going to school because schools are far away or are unaffordable. Similarly, a household is deprived if health facilities are too far away or lack enough staff to serve new clients.

Although the MPI was supposed to be updated every two years, consistent with the frequency of the district-level representative PSLM survey, the only measure of MPI currently available is based on the 2014-15 survey. Households deprived on 33% of the weighted indicators are categorized as poor. Table 2 shows the concordance between MPI and SPS. Households. While the magnitude of incidence is very different across the two measures, the ranking of the districts is very highly correlated (linear correlation coefficient of 0.74).

Table 2: Concordance between MPI (2014-15) and Simple Poverty Score (2015/16/17)

District	MPI Incidence	SPS Headcount	Median SPS	Mean SPS
Larkana	42.0	28.8	25.0	25.9
Dadu	51.4	40.9	21.0	22.5
Khairpur	51.6	28.7	24.0	25.9
Jamshoro	55.6	32.5	23.0	24.8
Matiari	62.1	28.6	25.0	26.2
Sanghar	66.8	34.9	23.0	24.0
Tando AY	67.3	31.3	23.0	25.3
Mirpurkhas	68.9	39.7	21.0	22.5
Kambar Shahdadkot	72.0	41.0	21.0	22.3
Badin	74.8	40.7	21.0	21.8
Tando MK	78.4	38.0	21.0	23.1
Thatta	78.5	40.1	21.0	22.2
Sujawal	82.0	46.9	19.0	20.6
Umerkot	84.7	44.9	19.0	20.9

Methodology for estimating rural poverty with ensemble transfer learning

The Social Protection Strategic Unit of Pakistan's Sindh Province has a goal of identifying households in rural Sindh eligible for cash relief, the kind support where households severely affected by Covid-19 and Monsoon floods are given financial assistance. Following these, we perform a binary classification task that predicts whether a household in a given 1km^2 grid cell is chronically poor using the median poverty score of the cell. Thus, each 1km^2 cell represents a single "training scene;" a model will learn from input about the scene from different data sources and predict the poverty status of the median household in that scene. For the poverty status variable, the PPRP report suggest the following categorization: 0 - 11 Extremely poor/destitute, 12 - 18 Chronically poor, 19 - 23 Transitory poor, 24 - 100 Non poor. We binarized these classes, categorising a cell with 0 - 18 PSC score as "chronically poor," and those with 19 or greater as "not chronically poor." Thus, we predict whether the median poverty score of a cell is below 19 (chronically poor) or 19 and above (not chronically poor). In machine learning classification tasks, there is often a decision to be made between optimizing for recall or precision. For a poverty classifier with good recall, most of the areas that are *truly* "chronically poor" would be predicted as chronically poor. In contrast, a classifier with good precision would instead focus on making sure that most of the predicted "chronically poor" areas are actually "chronically poor." The SPSU's cash transfer program, as with many poverty interventions, seek to ensure that everyone who needs support gets support, which drove us to optimize recall accuracy over precision in cases of trade-off between the two.

Through the SPSU, we accessed the SPS data containing 1.95 million households in 14 districts[2] of Sindh. While the data were meant to represent exclusively rural households, visual inspection revealed that some surveyed areas were *de facto* urban in nature, usually peri-urban settlements on the edge of medium-sized cities. Consultation with local experts in Pakistan confirmed that these areas were surveyed because their administrative status was 'rural' at the

[2] District boundaries in Sindh have seen considerable changes since the early 2000s, when there were 21 districts. 4 new districts were added in 2004, three in 2005, 1 in 2013 and 1 in 2020. We do not have access to the most recent boundaries and this paper uses boundaries consistent with districts in 2016. The 1.9 million geocoded household surveys are from 12 of the then existing 29 districts.

time. In line with our objective to map poverty in rural areas, we dropped these *de facto* urban observations. We also dropped all observations falling within or intersecting the boundaries of an "urban centre" as defined by the Global Human Settlement Layer (GHSL). Specifically, we used data from the 2019 Settlement Model (SMOD) of the GHSL to extract the urban centre layer (Pesaresi et al., 2019). We identified and dropped 95,271 observations falling within this urban centre layer. We also dropped 183,656 observations with (1) poverty scores exactly equal to zero – understood to be errors – or (2) GPS locational accuracy error over 20 meters, and left out 5,531 surveys conducted before 2016. Our final sample contains 1.67 million individual georeferenced observations.

Two types of spatial errors became evident while cleaning the data: (1) spatially diffuse GPS coordinates for individual settlements (often in fields or on roads), suggesting that the coordinates were not captured at the actual location of the household/settlement, and (2) unrealistically dense concentrations of observations in towns and cities from enumeration areas, suggesting that enumerators may have congregated at a location to upload data and accidentally assigned that location to all surveys that had been collected that day..

To create a target layer for the CNN model, we computed the median PSC score for observations falling within each 1km^2 grid cell. Given the spatial errors described above, we opted for a spatial resolution of 1km^2, substantially finer than the resolution used in the studies discussed earlier. This left 35,730 cells across Sindh that contained PSC observation(s). To preserve the anonymity of households, we eliminated 3,120 cells with fewer than 3 observations. We also dropped 329 cells with 300 or more observations, which probably reflected new urban settlements. Table 3 shows the descriptive statistics of median poverty scores for ~ 1.67 million households and 1km^2 grid cells.

Table 3. Descriptive statistics of PSC observations and target layer (1km^2 grid cells)

	Min	Max	Mean	Std	1st Quartile	Median	3rd Quartile
Household count (1km^2 grid cells)	3	299	47	48	14	31	62
Median poverty score (1km^2 grid cells)	3.5	80	21	5.5	18	21	24
Poverty scores (not gridded)	1	100	23	11	15	22	30

Input Data

We used three openly accessible inputs: daytime satellite imagery, nighttime satellite imagery, and accessibility data. Previous studies have shown that daytime satellite imagery and NTL offer important information about the economic geography of areas (Head et al., 2017; Jean et al., 2016; Bleakley and Lin, 2017). For data on daytime satellite imagery, we accessed 10m^2 resolution Sentinel 2 images from ESA's Copernicus Open Access Hub via QGIS. The images were captured between January and April 2016, contemporaneous with the SPS data collection period. All tiles except 1 had less than 1 percent cloud cover. The tiles were processed into true colour images and mosaicked into a single raster. For NTL data, we used the 2016 median VIIRS Annual VNL V2 product from the Earth Observation Group (Elvidge et al., 2021). The original resolution of the VIIRS image was ~ 500m^2, so we resampled it down to 10m^2 to match the Sentinel 2 resolution. Finally, we used as accessibility layer reflecting travel time to settlements with 5,000 – 10,000 population, which was extracted from global accessibility map (Nelson et al., 2019).

Ensemble transfer learning with convolutional neural networks

To predict whether the median household in each cell is chronically poor we employed an ensemble approach involving transfer learning among three models: (1) ResNet-50 (2) ResNet-50V2 and (3) ResNet-101. Past studies have used transfer learning techniques to map poverty and economic wellbeing with good results (Jean et al., 2016; Xie et al., 2016; Head et al., 2017). These studies employ a two-step approach where an existing CNN model is first used to predict nighttime light intensity using daytime satellite imagery as input. In the process, the CNN learns to extract predictive features from the daytime satellite imagery which are subsequently used as inputs in a regression to predict the final target label. We followed the approach of Yeh et al. (2020) by training the CNN models end-to-end using the three inputs (daytime and nighttime satellite and accessibility).

The three ResNet architectures were chosen for their high performance on the ImageNet image classification challenge. We predicted the poverty status for each cell using the three models in turns. In cases where the three models do not agree on a prediction for a cell, we used the majority prediction the final predicted status. All the models have three inputs (Sentinel 2 images, nightlights, and accessibility), and the data within each cell is at a 100x100 resolution. The models were initialized with weights trained on ImageNet.

For a given model and input, we extracted features with the corresponding ResNet architecture, performed global average pooling to reduce the extracted features, and added dense layers for classification. We then concatenated the final layers of each trained input and added a fully connected layer, which outputs a binary classification for each cell: chronically poor or not chronically poor. To minimize overfitting, we introduced dropouts prior to the final layer that randomly eliminated neurons.

The models were trained using ADAM to optimize the overall accuracy. A batch size of 16 and learning rate of 0.00005 were used in the training for all the models. The models were trained over 30 epochs with an early stopping mechanism that allows the models to stop after 10 continuous epochs if there is no improvement in the validation accuracy. The weights from the best performing epochs were retained. We implemented and trained the models using the Keras and Tensorflow libraries in Python.

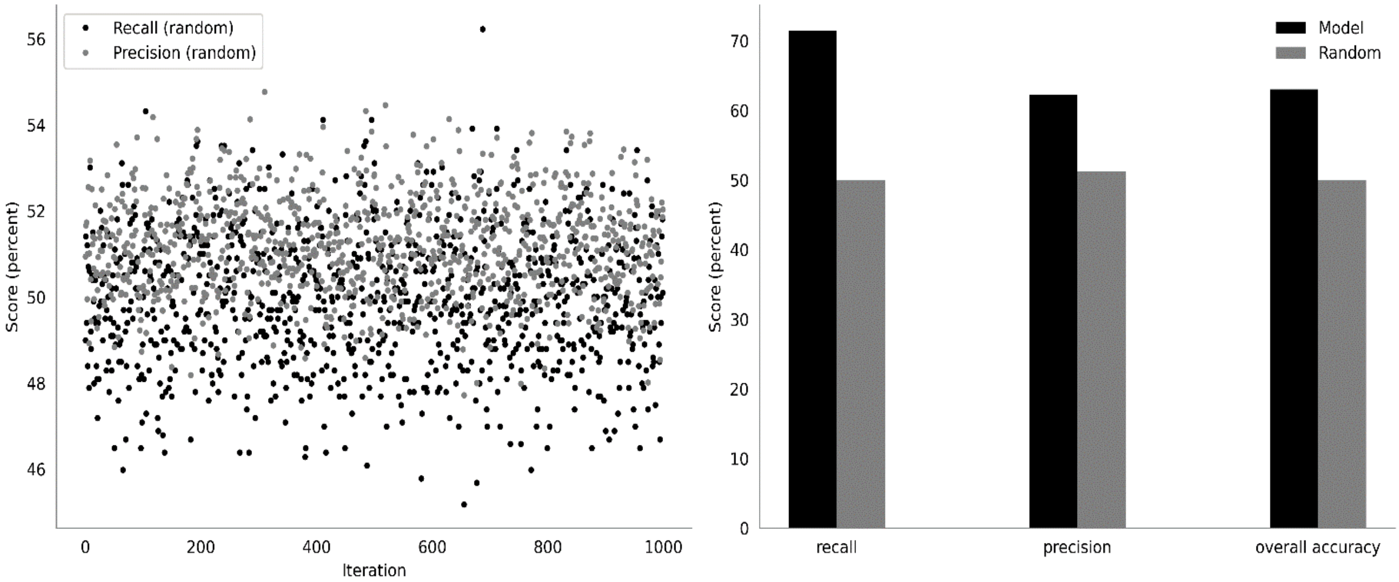
Results and Discussion

We validated the ensemble model using three approaches. First, we compared predictions for random holdout test samples. Second, we used a spatial cross-validation approach: whole districts were omitted from the training process and then used for out-of-sample testing. Last, we generated predictions for Ghotki, a district with no PSC data, conducted an original survey involving 7194 households sampled from 174 1km^2 grid cells, and compared our predictions with the survey data.

For the first approach, 30 percent of the cells across Sindh had PSC scores < 19, and 70 percent ≥ 19. To avoid bias in the training datasets, we sampled ~ 9600 cells (the count of cells with PSC < 19) from the latter so we have the same number of samples for each class. The dataset was split into 15,596 training, 1,732 validation, and 1,925 test samples. The model's performance in predicting the labels for the unseen test set is shown in Figure 2. To establish that the performance of the model is significantly better than a random occurrence or lottery, we generated random predictions for the test set, repeated 1000 times, and compared the results with the model.

In identifying the chronically poor, the model records 71 percent recall accuracy and 62 percent precision accuracy. The accuracy of the random predictions ranges from 44 to 56 percent for recall, and 47 to 55 percent for precision. As shown in Figure 2, the model performs 21 points better than the median of the random predictions for recall, and 12 points better for precision. Similarly, the model performs 13 points better than random predictions for the overall accuracy metric. These metrics highlight the importance of the model, especially in the context of Sindh where there is currently no measurement for mapping and supporting the chronically poor.

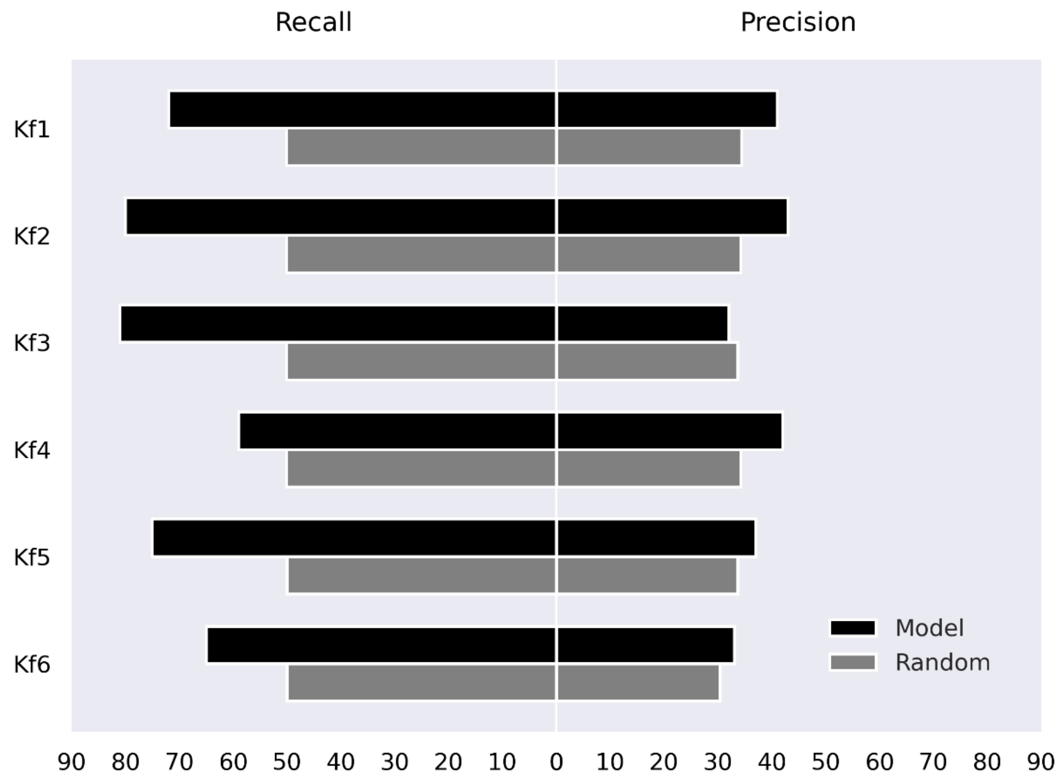
Figure 2: CNN ensemble model predictions versus random predictions for chronically poor



For a second validation approach, we randomly split the datasets into six folds (denoted "Kf"s), with each fold containing two districts. Then, we ran the ensemble model on five folds (ten randomly-selected districts) and tested the model's out-of-sample accuracy for one fold (two districts). The folds were rotated for each iteration such that all districts were used for both training and testing at the end of sixth round. This approach paints a better picture of the likely performance of the model when used to generate predictions in districts with no poverty data. As with the first approach, we compare the results of each cross-validation iteration with random predictions. We produced random predictions for the test sample, repeated 1000 times for each iteration, and the median score for recall and precision were computed.

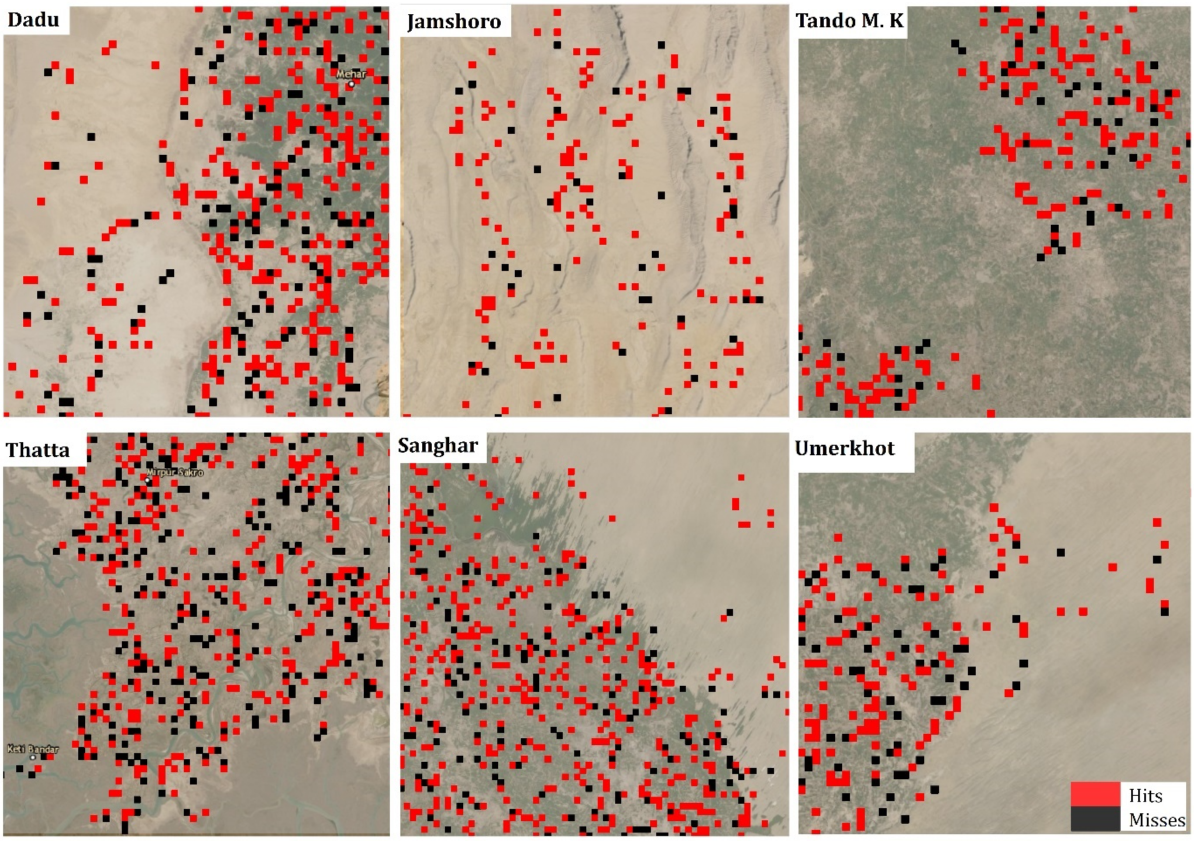
Figure 3 compares the performance of the model with the random predictions for each iteration. The model's recall accuracy ranges from 59 percent for Kf4 to 81 percent for Kf3 and a median of 74 percent. The model recorded at least 70 percent recall accuracy for four out of six folds. The precision accuracy ranges from 32 percent for Kf3 to 43 percent for Kf2. The model's recall accuracy is higher than that of the random predictions for all the folds. The precision of the model is also higher than the random prediction in all except one fold (Kf3). Even so, the model's marginal loss in precision is compensated by its recall performance, which is much higher than that of the random prediction.

Figure 3: Cross-validation comparison of CNN model with random predictions for the chronically poor.



The model’s recall performance is high in both arid and non-arid ecological regions of the province. The cross-validation performance of the model in selected districts is shown in Figure 4. The highest recall (81 percent) is observed in Kf3, which used predominantly non-arid districts (Tando M. K and Tando A. Y) as out-of-sample test districts. The second highest recall (80 percent), however, is found in Kf2 with test districts comprising of Matiari and the arid Jamshoro. The model’s recall is also high in test districts that have both arid and non-arid zones. For instance, the model recorded 75 percent recall accuracy in Sanghar and Mirphurkhas (Kf6), test districts with both ecological zones. Thus, the CNN model produces generally good results across ecological contexts, which is important because it shows the model is not biased against a particular ecological zone.

Fig 4: Cross-validation performance of CNN model in selected districts

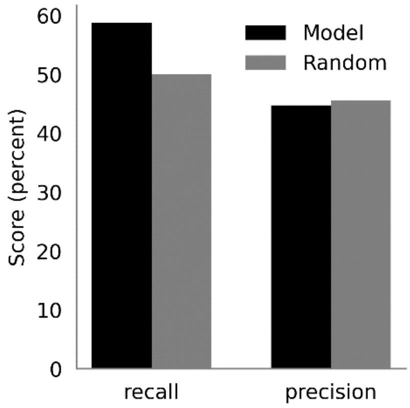


Hits: observed poor and predicted poor
Misses: observed poor and predicted not poor

Validation of existing CNN applications are mainly limited to the kinds of cross-validation approaches like those described above. We choose to go a step further and conduct a true out-of-sample validation exercise using an original survey in Ghotki, a district with no PSC data and both arid and non-arid ecological conditions. Prior to the survey, we used the CNN model to generate predictions for all habitable 1km^2 grid cells of Ghotki. Using the PSC methodology, we constructed poverty scores from the original survey data and used it as a benchmark to assess the performance of the model. As with the first two validation approaches, we also executed thousands of random predictions and compared it with the CNN model as shown in Figure 5.

The model's performance in the key recall metric is 59 percent while that of the random predictions (median) is 50 percent. The two are virtually at par for precision metric: 44 percent for the model and 45 percent for the random predictions. The model performing nine points better than random in the recall metric is significant considering that the model was trained on 2016-2019 PSC data while the original survey was conducted in 2022. Besides, unlike the original survey used for the validation, the PSC data used to train the model contain substantial spatial errors as described earlier.

Figure 5: Ground truthing results of CNN model versus random predictions



The ensemble deep learning model has produced promising results as shown above. In addition to outperforming all of the randomly generated predictions, the model also compares well with results from existing studies. For instance, the transfer learning model developed by Xie et al. (2016) identifies the poor in Uganda with 66 percent recall accuracy at 39 percent precision. This means in identifying the poor, results from holdout test samples randomly drawn across Sindh indicates that our CNN model performs considerably better than Xie et al.'s: five points higher in recall, and 23 points higher in precision. Thus, the model outperforms Xie et al.'s in minimizing both inclusion and exclusion errors in identifying the poor. This is even more significant considering the spatial resolution our model is 10 times finer than that of Xie et al.'s, which is 10km^2. Besides, Xie et al.'s poverty mapping includes urban and rural areas, making it less challenging than differentiating economic characteristics within rural areas as done here.

Similarly, the model's performance from the rigorous cross-validation is higher than Xie et al. (2016) in recall accuracy in four of the six iterations. The model's recall is at least 14 points higher than Xie et al.'s for two of the iterations (Kf2 and Kf3), and not less than six points higher for another two (Kf1 and Kf5). The recall of the model is lower than Xie et al.'s for only one iteration. The model's precision is higher than Xie et al.'s for three iterations (Kf1, Kf2 and Kf4) and lower in the other three. A potential reason for our model's comparatively average precision performance is inclusion errors in the underlying SPS. As shown in Figure 1, poverty scores below 19 often reflect households that are both asset-poor and consumption-poor, but there are also occasionally some asset-poor households that are not consumption-poor. Our target measure (the poverty scorecard) is chiefly asset-based, making it extremely challenging to identify these consumption-poor households when training on asset-poor scores from the sky. Besides, as indicated earlier, in cases of trade-offs, our priority to minimize exclusion errors in identifying the chronically poor than place higher emphasis on recall.

Conclusion

The traditional approach of using census to generate economic data is too expensive for LMICs, and such data are mostly out of date in rapidly growing developing countries. The Demographic and Health Surveys (DHS), the dominant alternative, has a limited sample and sparse spatial coverage in many countries. Capitalizing on advancement in computer vision techniques, there have been many new approaches that use deep learning techniques such as CNN in combination with satellite imagery to measure economic wellbeing or map poverty. However, CNN models are rarely developed to differentiate poverty within rural areas, and most prior applications have been too coarse in spatial resolution to use for fine-grained social support programs. Therefore, policy makers in LMICs seeking to target livelihood interventions in rural areas at a much finer scale will have major challenges relying on CNN models.

We have developed an ensemble CNN model based on three transfer learning sub-models to map chronic poverty at a fine scale (1km^2 resolution) in rural Sindh, Pakistan. The model draws on Pakistan's comprehensive but spatially noisy poverty scores data as target layer, and satellite imageries (daytime and night-time) and accessibility as input data. We have demonstrated that the combination of CNNs trained on publicly available inputs can generate good prediction of poverty at a much finer scale in rural areas, even when the target data is noisy. A rigorous cross-validation and external validation—ground truthing of predictions with an original survey—show that the model performs well in minimizing exclusion errors across both arid and non-arid regions, which are important in determining livelihood and lifestyle patterns in rural Pakistan. Altogether, our low cost and scalable approach to predicting rural poverty can improve how social welfare interventions are targeted in data challenged LMICs. Our approach's high prediction accuracy will also improve as less noisy data is collected in the future, underscoring the need for a more spatially-accurate economic datasets in LMICs that analysts can use to study and support social welfare interventions.

References

Atkinson, T., Cantillon, B., Marlier, E., & Nolan, B. (2002). *Social indicators: The EU and social inclusion*. Oup Oxford.

Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A., & Swartz, T. (2017). Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico. *arXiv preprint arXiv:1711.06323*.

Bleakley, H., & Lin, J. (2012). Portage and path dependence. *The quarterly journal of economics*, *127*(2), 587-644.

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, *108*(21), 8589-8594.

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, *119*(3).

Elvidge, C. D., Zhizhin, M., Ghosh, T., Hsu, F. C., & Taneja, J. (2021). Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. Remote Sensing, 13(5), 922.

Engstrom, R., Hersh, J. S. & Newhouse, D. L. (2017). Poverty from space: using high resolution satellite imagery for estimating economic well-being. *World Bank Policy Research. Working. Paper. 1,* 1–36.

Flechtner, S. (2021). Poverty Research and its Discontents: Review and Discussion of Issues Raised in Dimensions of Poverty. Measurement, Epistemic Injustices and Social Activism (Beck, V., H. Hahn, and R. Lepenies eds., Springer, Cham, 2020). *Review of Income and Wealth*, *67*(2), 530-544.

Filmer, D., & Scott, K. (2012). Assessing asset indices. *Demography*, *49*(1), 359-392.

Goldblatt, R., Heilmann, K., & Vaizman, Y. (2020). Can medium-resolution satellite imagery measure economic activity at small geographies? Evidence from Landsat in Vietnam. *The World Bank Economic Review*, *34*(3), 635-653.

Government of Pakistan (2016). National Poverty Report. Ministry of Planning Development and Reform, Islamabad.

Grosh, M., & Baker, J. L. (1995). Proxy means tests for targeting social programs. *Living standards measurement study working paper*, *118*, 1-49.

Head, A., Manguin, M., Tran, N., & Blumenstock, J. E. (2017, November). Can human development be measured with satellite imagery? In *Ictd* (pp. 8-1).

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American economic review*, *102*(2), 994-1028.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

Jerven, M. (2014). Benefits and costs of the data for development targets for the post-2015 development agenda. *Data for Development Assessment Paper*, *16*(9), 14.

Kuffer, M., & Barrosb, J. (2011). Urban morphology of unplanned settlements: The use of spatial metrics in VHR remotely sensed images. *Procedia Environmental Sciences*, *7*, 152-157.

Ladejinsky, W. (1969). Ironies of India's green revolution. *Foreign Aff.*, *48*, 758.

Lucci, P., Bhatkal, T., & Khan, A. (2018). Are we underestimating urban poverty? *World Development*, *103*, 297–310. https://doi.org/10.1016/j.worlddev.2017.10.022

Mellander, C., Lobo, J., Stolarick, K., & Matheson, Z. (2015). Night-time light data: A good proxy measure for economic activity? *PloS one*, *10*(10), e0139779.

Nelson, A., Weiss, D. J., van Etten, J., Cattaneo, A., McMenomy, T. S., & Koo, J. (2019). A suite of global accessibility indicators. *Scientific data*, *6*(1), 1-9.

Onda, K., Sinha, P., Gaughan, A. E., Stevens, F. R., & Kaza, N. (2019). Missing millions: undercounting urbanization in India. *Population and Environment*, *41*(2), 126–150. https://doi.org/10.1007/s11111-019-00329-2.

Pesaresi, M., Florczyk, A., Schiavina, M., Melchiorri, M.,; Maffenini, L., (2019). GHS settlement grid, updated and refined REGIO model 2014 in application to GHS-BUILT R2018A and GHS-POP R2019A, multitemporal (1975-1990-2000-2015), R2019A.

Persello, C., & Stein, A. (2017). Deep fully convolutional networks for the detection of informal settlements in VHR images. *IEEE geoscience and remote sensing letters*, *14*(12), 2325-2329.

Perez, A., Yeh, C., Azzari, G., Burke, M., Lobell, D., & Ermon, S. (2017). Poverty prediction with public landsat 7 satellite imagery and machine learning. *arXiv preprint arXiv:1711.03654*.

Sahn, D. E., & Stifel, D. (2003). Exploring alternative measures of welfare in the absence of expenditure data. *Review of income and wealth*, *49*(4), 463-489.

Sebastian, A. R., Shivakumaran, S., Silwal, A. R., Newhouse, D. L., Walker, T. F., & Yoshida, N. (2018). A proxy means test for Sri Lanka. *World Bank Policy Research Working Paper*, (8605).

Schreiner, M. (2006). Simple Poverty Scorecard Poverty-Assessment Tool: Bangladesh. *SimplePovertyScorecard. com/BGD_2000_ENG. pdf, retrieved September*, *30*, 2008.

Schreiner, M. (2014). How do the poverty scorecard and the PAT differ? *microfinance. com/English/Papers/Scorecard_versus_PAT. pdf, retrieved*, *4*.

Sen, A. K. 1985. Commodities and capabilities. North-Holland, Amsterdam.

Serajuddin, U., Uematsu, H., Wieser, C., Yoshida, N., & Dabalen, A. (2015). Data deprivation: another deprivation to end. *World Bank policy research working paper*, (7252).

Skoufias, E., Diamond, A., Vinha, K., Gill, M., & Dellepiane, M. R. (2020). Estimating poverty rates in subnational populations of interest: An assessment of the Simple Poverty Scorecard. *World Development*, *129*, 104887.

Sun, J., Di, L., Sun, Z., Wang, J., & Wu, Y. (2020). Estimation of GDP using deep learning with NPP-VIIRS Imagery and land cover data at the county level in CONUS. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 1400-1415.

Tapiador, F. J., Avelar, S., Tavares-Corrêa, C., & Zah, R. (2011). Deriving fine-scale socioeconomic information of urban areas using very high-resolution satellite imagery. *International journal of remote sensing*, *32*(21), 6437-6456.

Wang, J., Kuffer, M., Roy, D., & Pfeffer, K. (2019). Deprivation pockets through the lens of convolutional neural networks. *Remote sensing of environment*, *234*, 111448.

Watmough, G. R., Marcinko, C. L., Sullivan, C., Tschirhart, K., Mutuo, P. K., Palm, C. A., & Svenning, J. C. (2019). Socioecologically informed use of remote sensing data to predict rural household poverty. *Proceedings of the National Academy of Sciences*, *116*(4), 1213-1218.

Wurm, M., & Taubenböck, H. (2018). Detecting social groups from space–Assessment of remote sensing-based mapped morphological slums using income data. *Remote Sensing Letters*, *9*(1), 41-50.

Xie, M., Jean, N., Burke, M., Lobell, D., & Ermon, S. (2016, March). Transfer learning from deep features for remote sensing and poverty mapping. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., ... & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications*, *11*(1), 1-11.