Using satellite imagery, transfer learning, and survey data to predict poverty

*†

Laniah Lewis [‡] Taymour Siddiqui [§] Michael Chan [¶]

Rus Adamovics-Davtian [‖]

September 3, 2021

Abstract

Accurate socio-economic measures are vital to the success of poverty-intervention programs yet are unavailable in regions of the world where they are most needed. Models that can predict poverty using satellite imagery and household survey data are, therefore, invaluable for helping households most in need. Using an existing two-part model, we aim to improve the accuracy of poverty predictions for an unconditional cash transfer program in [redacted]. By altering several aspects of the model including the range of spectral bands for the images used to predict nighttime lights, and optimizing the hyperparameters of both stages of our model, we are able to increase the accuracy of poverty predictions by 3.8 percent.

Keywords: Asia, Poverty, Cash Grants, Satellite Images, Transfer Learning, Deep Learning.

1 Introduction

More than 9.2 percent of the world, or 689 million people, lives in extreme poverty on less than 1.9 dollars per day (The World Bank, 2020). Fortunately, over the past few decades, poverty has continued to decline, reducing 1 percent every year from 1990 to 2015.

*This paper has been censored & edited in order to uphold the standards and personal relations of the World Bank and its members. This work is intended for academic use only and should be treated in such manner.

†Thank you to Professor Katya Vasilaky for your technical guidance throughout this project as well as your insightful revisions to the final paper. Thank you to World Bank team members Rob Marty and Alice Duhaut for allowing us to join your team and being great mentors. Thank you to the California Polytechnic DxHub for connecting us with the World Bank and providing guidance and stellar technical support.

‡Cal Poly, Department of Economics (e-mail: lilewis@stanford.edu)

§Cal Poly, Department of Economics (e-mail: tasiddiq@calpoly.edu)

¶Cal Poly, Department of Economics (e-mail: mchan92@calpoly.edu)

‖Cal Poly, Department of Economics (e-mail: radamovi@calpoly.edu)

However, the COVID-19 pandemic has impeded this progress. The World Bank's 2020 report on poverty projects that over 150 million people could fall back into extreme poverty as a result of the pandemic (The World Bank, 2020).

A huge challenge is figuring out which areas and households are most affected by the pandemic and are struggling at higher proportions. This allows organizations fighting poverty to effectively use their resources; however, measuring poverty has always been challenging, even before the pandemic. Surveys/censuses are a traditional method of measuring poverty but are not conducted often enough to adapt to new issues like COVID-19 and are very costly. Over the past decade, finding alternative, cheaper ways of measuring poverty and other economic indicators has been an incredibly important area of research.

In this paper, we use a satellite imagery approach to measuring poverty because it is cheap and it makes it possible to get more updated measures of poverty than other methods. We will use deep learning Convolutional Neural Networks (CNNs) to predict poverty using satellite imagery. We will be working with an earlier CNN model provided by the World Bank and will make 3 improvements to achieve more accurate predictions that are helpful in targeting poverty. First, we introduce multispectral imagery into our model to incorporate more information that our model can learn from. Secondly, we improve the model from achieving just a binary measure of poverty to a continuous measure of poverty to differentiate between different poverty levels. Finally, on top of measuring poverty, we alter the model to allow us to also predict other economic measures of well-being.

Many policy strategies have been taken to reduce worldwide poverty. The earliest programs focused on conditional cash transfer programs (CCTs) which give money to poor people in return for meeting certain conditions. For example, the PROGRESA program disbursed cash transfers to households in Mexico if the individuals within recipient households engaged in specific behaviors regarding medical care and health education (Gertler and Boyce, 2001). CCTs have had some success in the past (**?**), but they require a large amount of bureaucratic oversight to be administered. An arguably more efficient alternative to conditional cash transfers, are direct cash transfers, which provide financial aid without any conditions. For example, GiveDirectly is giving monthly payments to impoverished villages in Kenya for a total of 12 years with no conditions attached (Aizenman, 2017). One of the main benefits to these direct cash transfers is the limited oversight required. Instead of administering a universal basic income for an entire country, cash transfers tend to target specific groups in financial need (Banerjee et al., 2019).

The effectiveness of cash transfer programs is constantly under scrutiny. The long-term effects of cash transfer programs are still unknown since most studies involving these programs have difficulties measuring long-term effects or were administered recently (Aizenman, 2017); however, there is evidence that cash transfers, conditional and unconditional, benefit the communities receiving aid in the short run. For example, a conditional cash transfer program in Malawi increased school attendance and decreased sexual activity in young women (Baird, McIntosh, and Ãzler, Baird et al.). The unconditional GiveDirectly cash program in Africa increased access to medication, access to household necessities, and allowed some individuals to save funds or start businesses (Aizenman, 2017). However, these cash transfer programs are only effective when impoverished groups are identified correctly. The positive impacts of cash transfer programs are diluted when groups are given aid based on affiliations (such as political or cultural) as opposed to financial need (Banerjee et al.,

2019).

[redacted] has found it challenging to determine which households to target. The aim of the program is to help bring people out of poverty rather than keeping them dependent on the support of this program. For this reason, it becomes important to be able measure changes in poverty and determine which households still need the support and those that don't. Furthermore, research suggests that different rates of disbursement among provinces may even be rooted in political causes rather than based off of need [redacted]. For this reason, having accurate poverty measures for [redacted] would help hone in on which areas should receive cash grants to maximize the impact of the cash grant program and reduce the cost of unnecessary transfers.

A traditional method for estimating poverty has been via door-to-door household surveys; however, this is expensive and time consuming (Tingzon et al., 2019). Burke et al. (2021) analyzed prior research methods in estimating poverty using satellite imagery and concluded the following about household surveys: "surveys are typically only representative at the national or (sometimes) regional level, meaning they often cannot be used to generate accurate summary statistics at a state, county, or more local level" (Burke et al., 2021). Our research focuses on estimating poverty levels at locations below the poverty line, which happen to be at local levels. One of the newer and cost effective approaches to measuring poverty is through satellite imagery. This approach can be traced back to research by Henderson et al. (2012), which describes how economic activity can be measured from space. The researchers introduced US Air Force Weather Service satellite night-lights data as a useful proxy for economic activity for regions with the poorest economic data quality, rated from A (best) to D (worst). "Almost all industrialized countries receive a grade of A. By contrast, for the 43 countries of sub-Saharan Africa, 17 get a D and 26 get a C" (Henderson et al., 2012). The researchers gave a greater focus to grade D countries, which include many African countries. They estimated true income growth from 1992 to 2003 by combining information on measured income growth with night-time light information through a fixed effects model and achieved a R^2 of 0.66. This analysis paved the way for further research attempts at improving the original methods of Henderson et al. (2012) in developing countries such as the Philippines and India (Fatehkia et al., 2020; Tingzon et al., 2019; Yeh et al., 2020).

There is plenty of literature that support the use of night-time satellite images of lights to estimate several measures of human well-being. While human well-being can be quantified in many different ways, night-time light images have been successfully used as important inputs for estimating figures like GDP, electricity access, and poverty in varying models. The methods that are used in estimating these measures of well-being start from simple correlation analysis and regression modelling and can be as complex as deep-learning algorithms (Ghosh et al., 2013). The wide variety of different approaches to using night-time lights as a proxy variable in models for human well-being indicate that remote-sensing data can serve as an important element in estimating poverty.

Recent research discussed by Burke et al. (2021) involve satellite imagery to predict poverty that goes beyond the original approaches of Henderson et al. (2012). These methods involve powerful machine learning neural network models that use satellite images as inputs to predict wealth indexes. With noisy and limited training data of villages and their corresponding poverty levels, researchers deployed more creative machine learning methods

including simulating synthetic data, transfer learning, and unsupervised learning. Despite the lack of quality training data, the model performance of these advanced methods remained highly stable and robust to various types of training noise. According to Burke et al. (2021), "information derived from satellites could always explain more than half, and often more than 75%, of the variation in the survey-measured asset wealth, with performance appearing to trend upward over time." This statement summarizes 12 different studies that used imagery in combination with other features to predict economic growth at local levels in the developing world. Some of the best performing models use high-resolution satellite imagery that are gathered from private APIs and are expensive, but others have accomplished comparable results with free, publicly available satellite imagery.

Privately-accessed satellite images can be costly, time consuming and difficult to access, so many researchers have attempted to use publicly available data, such as Landsat 7 satellite imagery (**?**). Other alternatives to privately-accessed satellite images include: OpenStreetMap Data (Tingzon et al., 2019), individual level data such as Facebook User Data (Tingzon et al., 2019) and mobile phone data (Steele et al., 2017). The combination of satellite and individual level data provides an opportunity for estimation of poverty at an individual or household level. Models that use a combination of phone usage data with satellite images in statistical methods have been able to achieve predictions down to the neighborhood or even household level rather than the larger regional or provincial sizes of estimation (Steele et al., 2017). This presents an opportunity for more targeted policy and action to be present, as it is more informative for a cash grant program to differentiate poor neighborhoods or households. There are different measures of wealth that can be used to estimate poverty with varying success. For example, Steele et al. (2017) found that predicting an asset-based wealth index performed better in their combined phone data and satellite image geostatistical model than models that attempted to estimate income or PPI. Though income, PPI, and asset-based wealth indexes are all ways to measure human well-being, their estimations may require different modeling approaches.

Some newer methods that estimate poverty using satellite imagery have also tried to measure changes in poverty, as opposed to just getting a measure of poverty at one point in time (Yeh et al., 2020). However, some researchers have struggled to adapt the satellite learning approach to recognizing changes in poverty and economic development. Kondmann and Zhu (2020) have tried in Rwanda using Landsat 7 data and found that the transfer learning strategy struggles to recognize changes (Kondmann and Zhu, 2020). This may be because it is difficult to spot economic changes through imagery alone because the economic improvements may be more subtle. Additionally, they used Landsat 7 images which have low resolution. In our research regarding [redacted], we are also using low-resolution Landsat images.

Our paper seeks to build on the research mentioned above. Much research has been done to measure poverty outside our country of interest, we seek to extend those methods to [redacted]. In addition to applying the transfer learning approach to [redacted], we aim to improve this model in 3 different ways: incorporating multi-spectral imagery into the model, predicting poverty as a continuous variable and using the model to predict other economic variables. These contributions to previous methods allow for a deeper evaluation of the effectiveness of social programs in [redacted]. Jean et al. (2016) and Yeh et al. (2020) have made similar contributions to this field of research while focusing on African countries.

We plan to compare our results to the results obtained by Jean et al. (2016) and Yeh et al. (2020). Therefore, our research can also be used to determine how generalizable previous methods are to countries outside of the ones that have already been studied.

2 Data

We make use of several datasets to estimate our convolutional neural network (CNN) model: Landsat 8 and VIIRS satellite images. The CNN is trained on larger image datasets to help accurately choose the features needed to predict poverty. Once the features are selected from the CNN, we use the OPM data, which contain our poverty outcome data. Below we detail each data source.

Daytime light intensity

We use Land Remote-Sensing Satellite System (Landsat) 8 images from 2014 from the Earth Engine Data Catalog. There are hundreds of thousands of daytime images but due to sampling methods for nighttime light discussed in the next section, we utilize 50,583 daytime satellite images from areas in [redacted]. Each image has a resolution of 48 by 48 where each pixel has 3 values corresponding to red, blue, and green pigments. Landsat images also have visible and near-infrared bands to assess brightness temperature. To prepare this data for the CNN, the daytime light values are outputted as arrays. This data was then extracted into different arrays organized around [redacted] coordinates and each array was given a unique identifier so that features could later be extracted from these arrays using the CNN.

The advantage of this data is that it has multi-spectral imagery but a limitation is that they have a low to medium resolution so this makes it harder to pick up on different features that we might be interested in, like roof material.

Nighttime light intensity

We use Visible Infrared Imaging Radiometer Suite (VIIRS) images from 2013-2018 using the Earth Engine Data Catalog. It contains monthly and annually averaged night time light radiance measures. Since our daytime images are from 2014, we filter out all other years of nighttime radiance values. We created a categorical and continuous measure of night time light radiance. The original night time radiance values are continuous and mostly located within the [0,1] range but outliers are heavily right-skewed; therefore, we add 1 and perform a log transformation. In order to build the categorical measure, we use k-means clustering to discretize each continuous value into 3 levels of night time light coded as 0 (low), 1 (medium), or 2 (high). There are over 1.8 million nighttime light observations but the 3 levels are highly unbalanced. In order to have perfect balance among the 3 classes, we down sample the two majority classes to the size of the minority class. Now, all 3 levels of nighttime light occur with $\frac{1}{3}$ frequency. Each class contains 16,861 nighttime light values for a total of 50,583 values. For each daytime image at a specific area in [redacted], we merged its corresponding level of categorical and continuous annual night time light. We do not use monthly night time radiance values due to having only one 2014 daytime image per specific area so that we may keep a one-to-one merge. To prepare this data for the CNN, the nighttime light labels

are outputted as an array and a parameter dictionary, such as the number of nighttime light labels, is created.

[redacted] Survey

[redacted] survey data is a proprietary dataset that includes survey data describing individuals in [redacted]. The de-identified [redacted] data and [redacted] coordinates were cleaned up, which appended relevant variables (such as asset and consumption data) and created anonymized household IDs and coordinates. At the household level, each observation of a surveyee appears multiple times over time creating a panel dataset structure with a total of [redacted] households in the year 2014. Each observation includes demographic data including age, gender, marriage status, and residency status. More importantly, each participant is labelled if they are a [redacted] recipient. A normalized [redacted] poverty score is included for each participant. Since the score is normalized, the threshold for eligibility for [redacted] benefits is 0. Geographic information- such as whether they live in an urban or rural area, or what province they live in- is also available.

The [redacted] survey data also includes answers to several questions that surveyors must ask or determine while conducting their surveys. Questions range from asking about details regarding a participant's education to whether or not a surveyor considers the surveyee to be disabled. An important fact about these survey questions is that not all questions are filled in for all participants in all periods. Moreover, it is quite common for some questions to not have an answer at all. This portion of the data may provide some interesting insight and detail about some participants, but has a risk of being up to interpretation to surveyors or missing too many responses to warrant usage in models, which may lead to bias in the data. Another important quality to note about the dataset is that not all households are present in all survey years. Table 1 shows data for the year 2014.

Table 1: Survey Data
Mean Statistics

Age Statistics Per HH	*Oldest Person*	*Youngest Person*	*HH Head*
	55.72	3.6	46.63
Localities	*Urban Proportion*	*Rural Proportion*	
	25.43%	74.57%	
Unconditional Cash Grant	*Proportion of Recipients*	*Proportion of Non-Recipients*	
	[redacted]%	[redacted]%	

HH = household.

3 Problem description

The aim of this paper is to improve the prediction accuracy rate given an existing model proposed by Marty and Duhaut (2021). The model uses a transfer learning approach with a pre-trained CNN model. The satellite images are taken as inputs into this CNN model to predict night-time lights (NTL) with day-time lights (DTL). This is referenced as the first-stage CNN model. After the first stage CNN model, the dimension of the relevant

features are first reduced via principal component analysis (PCA). The resulting features left from the PCA dimension reduction are extracted and are then used to train machine learning models that will predict a binary asset-based poverty measure.

The reason for first training a CNN model of NTL predicted by DTL is to determine the most relevant features of DTL in predicting poverty. Since, the data from DTL are large, our aim is to reduce the dimensions of these images before using them to predict an actual poverty index, which is a far smaller dataset. We will extract the most relevant features that the CNN model finds in the training process and transfer these features to a related problem in predicting poverty. We then merge these extracted features with the OPM data and use a grid-search method consisting of several statistical and machine-learning models to find the most accurate model that predicts a binary asset-based poverty measure. These models include: Linear Support Vector Classification, Decision Tree Classification, Bagging Classification, Gradient Boosting Classification, Random Forest Classification, ADA Boost Classification, K-Neighbors Classification, and Gaussian naive Bayes.

The initial accuracy of the first stage CNN model is 0.69 in the year 2014, which means that 69% of the NTL predictions from DTL satellite images were correct in categorizing a household's level of poverty. The initial accuracy of the machine learning model used to predict the binary asset-based poverty measure from OPM data merged with the first stage CNN features was .449. This means that the original model predicted poverty correctly 44.9% of the time. We aim to improve this model in 3 different ways: incorporating more spectral bands, tuning the hyperparameters of the pre-trained CNN model predicting night-time lights, and optimizing the model selection grid search for binary and continuous output.

The inputs of the current model only take in three main visible spectral bands that include red, green, and blue. Spectral bands are captured image data that represent a specific wavelength range. However, near-infrared bands that human eyes normally cannot observe from the LANDSAT satellite images are also available and may have important features that a CNN may be able to extract to use in predicting poverty. Our first improvement to the model is incorporating more spectral bands. Each satellite image that we have has 7 different bands, where each band captures light from different parts of the electromagnetic spectrum. The importance of incorporating different bands is that they can all provide unique features that would be helpful in predicting poverty. Different combinations of bands are helpful in measuring different things. For example, a combination of the Near Infrared and the Red spectral bands is very good at picking up vegetation in satellite images. Our goal is to incorporate all the bands so that we can capture all the important features we have available. The difficulty in including all 7 bands is that the pre-trained Convolutional Neural Networks (CNN) model that we use has been pre-trained on millions of RGB images and is designed to only take up to 3 bands as input. We could train our own CNNs but this is time consuming and data-intensive. Helber et al. (2017) found that inputting non-RGB bands individually into these pre-trained CNNs still maintains fairly high classification accuracy so we should be able to input all of the bands individually into our pre-trained CNN and still get accurate results. So we will estimate a total of 5 models, one for the RGB values and 4 models for each of the rest of the 4 bands being plugged into the CNN model individually. After estimating these models, we will extract the features from each one, combine them and use those features to make predictions about poverty.

We will also tune the hyperparameters of the pre-trained CNN model to improve

the overall model accuracy. By changing the hyperparameters that are not learned by the CNN model we seek to improve the performance of the predictions of the CNN model. Changing the hyperparameters of the CNN will be an important part of the process of improving our overall model. These hyperparameters are parameters that we provide the CNN to control different parts of the learning process. While some machine learning models may not have many hyperparameters, convolutional neural networks have several that affect model performance. The initial RGB bands model extracts the last layer of a pre-trained model named VGG16 and is fed into 1 dense hidden layer of 100 nodes with a Rectified Linear Unit (ReLU) activation function and drop out rate of 0.3 to avoid overfitting. This connects to the output layer where there are 3 nodes for our 3 night-time lights classes with a softmax activation function- a type of generalized logistic function- to output probabilities of whether each daytime image of land is associated with low, medium, or high nighttime light radiance. This last output layer has 3 nodes, which is equal to the number of bins that separate NTL values into 3 separate classes. Potential hyperparameter changes include: making changes to the number of nodes within the dense layer, adding convolutional layers, adding more hidden dense layers, changing the number of epochs, changing the batch size, and changing early stoppage occurrence. Batch size here is a reference to the number of observations within the training set. We hope that these adjustments will improve model performance.

Finally, we aim to optimize model selection by expanding our machine learning gridsearch models over a larger hyperparameter space. The best model of the original gridsearch gives an accuracy score of approximately .449. This means that the original model predicted poverty correctly 44.9% of the time. Having a more extensive gridsearch will allow us to predict poverty with better perfomance metrics. We plan to accomplish this by: including more parameters within the classifier models and expanding the ranges associated with those model parameters. Although this may be time consuming, it is worth pursuing to increase our accuracy, precision, recall, and F1 scores. In addition, the original gridsearch is only prepared for a binary outcome, namely whether a household was in poverty or not. Expanding the gridsearch to handle continuous outputs will allow us to observe where a household falls on the financial health spectrum. Expanding the gridsearch for continuous outputs requires us to add more regressor models including Lasso and ElasticNet, more model parameters, and larger ranges for model parameters. The criterion for the continuous gridsearch will be mean squared error (MSE), correlation, and R^2 since we will be evaluating a regression model with continuous output. Once we have tuned the gridsearch at the binary and continuous levels, we will test how well our model can predict asset indices and individual assets to evaluate if they are viable proxies for poverty.

4 Results

In this section we establish our main results where our new model helps us achieve higher accuracy, achieve a continuous measure of poverty and also additionally estimate an asset index using our model. We start off in our first subsection, with a description of the specific changes we made to the CNN model to achieve a higher accuracy. In our second subsection, Binary and Continuous Poverty Estimation, we show how we increase accuracy

for the binary measure of poverty and achieve a continuous poverty measure. In our third subsection, we include some additional findings including how we were able to use the model to evaluate asset indices as proxies for poverty. In our final subsection, we incorporate a robustness check of the CNN model.

4.1 NTL Predictions Using DTL Data

Table 2 below shows the changes that we made to the first stage of the model to achieve better accuracy. Overall the model is similar to the original; however, we made 3 key changes that provided us higher accuracy. First, we expanded the number of bands in our model from 3 (only RGB) to 7 bands including the different types of multispectral imagery providing more features for the model to learn from. Second, we reduced our batch size which generally results in more efficient models that run more quickly and have higher accuracy. Finally, we reduced the early stoppage occurrence. All of these changes, helped us make some improvement in the overall accuracy of the model as can be seen in Table 3 of the next sub-section.

Now that all 7 bands have been included, predictions for NTL values include single band images. The pretrained model that the original RGB CNN model uses accepts only three bands as an input. The workaround used to get single band image data working with this model was to repeat the translated image data three times to mimic the shape of an RGB array. The alternative is to create a CNN model from scratch which would sacrifice the advantage of using a pre-trained model that has been trained with a large dataset. The repeated single band data is not an RGB array and the effects of using single band data this way on accuracy of prediction are not clear to us.

Table 2: First Stage Model Characteristics
Prior To & Following Contributions

	Prior	*Following*
Bands	3	7
Batch Size	500	32
Early Stoppage Occurrence	10	2

First Stage CNN predicts NTL from DTL

4.2 Binary and Continuous Poverty Estimation

After extracting the features from the updated First Stage model, we ran the updated grid search for poverty on a binary scale. As shown in Table 3, we increased the accuracy and precision scores but decreased the recall and F-1 scores. Our predictions could potentially be used by the [redacted] to determine which households to follow up with. These predictions may also be used by the World Bank to guide the selection of eligible households. Therefore, increasing accuracy is vital for the World Bank to allocate resources efficiently and sufficiently to households in need. Our contributions were able to increase the the accuracy score by .038 (from .449 to .487). In terms of accuracy, we are predicting poverty at the binary level 3.8% more accurately than the original model. Though this may seem

like a small increase, this could be quite a large number of households depending on sample size. Given the nature of how these predictions will be used, any improvements should be applauded. Overall, we are currently predicting the binary level of poverty correctly 48.7% of the time. This is close to as good as a coin flip but not as accurate as a researcher would want due to the influence these predictions will have. Therefore, this approach is not 100% foolproof.

The original gridsearch was limited to handling poverty through a binary perspective: either a household was in poverty or they were not. We expanded the gridsearch to a continuous scale to evaluate where households land on the spectrum of financial health. We implement the following machine learning models: Ridge, Lasso, ElasticNet, Linear Support Vector Regressor, Decision Tree Regressor, Bagging Regressor, Gradient Boosting Regressor, Random Forest Regressor, ADA Boost Regressor, and K-Neighbors Regressor. In addition, we use performance metrics such as R^2, mean squared error (MSE), and correlation. We achieved an R^2 of .0319, an MSE of 134.7657, and a correlation of .18174. Because R^2 is one of the most common goodness-of-fit measures, we analyzed this closely. According to the R^2, the model only explains 3.19% of the variance from the independent variables that predict poverty. This is quite low, though it does not mean that our model is useless for continuous measures of poverty. It's possible that the nature of our data has a lot of unexplainable variability, which causes low R^2. However, this low R^2 paired with such a high MSE and low correlation shows that our model is not the best for predicting continuous levels of poverty. This aspect of the model could definitely be improved upon and it may be helpful to add another criterion such as the Akaike Information Criterion (AIC) to double-check for over-fitting, though this doesn't seem to be an issue currently.

The very basis of our approach has already been applied to countries such as the Philippines, Rwanda, and Bangladesh (Kondmann and Zhu, 2020; Steele et al., 2017; Tingzon et al., 2019). We expanded these methods by incorporating more spectral bands, tuning the hyperparameters of the pre-trained CNN model predicting night-time lights, and optimizing the model selection grid search for binary and continuous output. These contributions will be externally valid because they are just adjustments of the original methods. It should be noted that our methods struggled to accurately predict poverty at a continuous level, which may or may not be externally valid for other countries. To increase external validity, the pre-trained CNN could be trained on more images, potentially from other regions of the world.

Table 3: Machine Learning Model Performance Metrics
Prior To & Following Contributions

Binary	*Prior*	*Following*
Accuracy	.449	.487
Precision	.452	.465
Recall	.964	.8922
F1	.616	.6121
Continuous		
R2	N/A	.0319
MSE	N/A	134.7657
Correlation	N/A	.18174

Machine learning model predicts poverty on a binary and continuous level

4.3 Asset Index Estimations

We performed PCA on the OPM asset data and extracted the first component to create an asset index. We used this asset index consisting of over 20 assets to create sub-indices for amenities, appliances, transportation, and entertainment. We used these indices as proxies for poverty. For example, a household that is well off will have a sanitation area and/or a refrigerator, so households that do not have those assets are likely to be living in poverty. We created a total of 10 asset indices using PCA:

1. Main Asset Index
2. Additive Main Asset Index
3. Amenity Asset Index
4. Additive Amenity Asset Index
5. Appliance Asset Index
6. Additive Appliance Asset Index
7. Transportation Asset Index
8. Additive Transportation Asset Index
9. Entertainment Asset Index
10. Additive Entertainment Asset Index

It should be noted that asset indices have limitations that may matter when using them as a proxy for poverty (Vyas and Kumaranayake, 2006). Asset indices frequently reflect long-run households that have acquired assets over time. This means that asset indices fail to account for short-run and/or temporary shocks that effect financial health, such as COVID19. Ownership of an asset does not account for the quality of an asset, which is directly related to poverty. For example, a family with a brand new shower will likely have more money than a family with a 30 year old shower that may have wear and tear. In addition, there exists inconsistencies across sub-groups on what an asset signifies. For example, in some regions having a bicycle is an indicator of wealth while in other regions a bicycle may be an indicator of poverty. Finally, the additive asset indices are limited because they give equal weight to assets that should not be equally weighted, such as a stove and a television.

In Figure 1 we plotted the correlation matrix between the asset indices and the

poverty score (pscore). We then evaluated which asset indices would be appropriate proxies for poverty. We can see that the additive asset (51%), asset (46%), and additive appliance (45%) indices are most heavily correlated with poverty. We then used the continuous grid search to see how well we could predict the 10 asset indices, though this paper only analyzes the relevant asset indices determined by the correlation matrix. Analyzing only R^2 from Table 4, the explanatory variables (CNN features) fit the data better for the asset index, additive asset index, and additive appliance index as opposed to the poverty score. Therefore, at a continuous level, the explanatory variables (CNN features) explain more variation in specific asset indices than the actual poverty score.

Table 4: Asset Indices Performance Metrics

	R2	*MSE*	*Correlation*
Main	.0774	.412	.2785
Additive Main	.0589	3.5188	.2511
Additive Appliance	.144	.4369	.395

Additionally, we used the binary grid search to predict whether a household has an individual asset or not. As can be seen in Table 5, the fan asset (including ceiling, table, pedestal, or exhaust fans) achieved an accuracy of 0.9126, precision of 0.9184, recall of 0.9918, and an F1 score of 0.9537. Therefore, it is possible that our model is able to predict whether a household will have a fan or not with higher accuracy, precision, recall, and F-1 scores than when we attempt to predict the poverty score. If we treat having a fan as a proxy for whether a household is in poverty or not, then we would be able to predict poverty with significantly better performance metrics than the original model. It should be noted that approximately 89.05% of households had fans. Therefore, approximately 10.95% of households did not have fans and are considered to have been in poverty. This is quite smaller than the 30.75% of households determined to be in poverty utilizing the [redacted] survey poverty score. Therefore, it is possible that the performance metrics from the fan asset are suffering from over-fitting. Besides the fan asset, the majority of individual assets had accuracies close to 1; however, the precision and recall scores dropped significantly towards 0. Given the nature of this research, precision and recall are both important so we will not discuss these assets as a potential proxy for poverty.

Table 5: Performance Metrics of Fan Asset

	Accuracy	*Precision*	*Recall*	*F1*
Fan Asset	0.9126	0.9184	0.9918	0.9537

4.4 Robustness Within Model

If this model were to have any changes, we want the model to be robust to the changes and have consistent outputs. These changes can include different hyperparameters. An example of hyperparameters that we changed to test for stable results was the batch size value. We used a value of 32 for the batch size hyperparameter in the first stage CNN

model. To check for robustness, we also used values like 64, 128, 256, 300, and 1000 when estimating the CNN model. There were no obvious changes in outputs with these different batch sizes. Stable prediction values are important for this early stage prediction process, since the following stage predictions depend on the features of this CNN model.

5 Discussion and Conclusion

Combating poverty has always been an important focus for governments, policy-makers, and researchers, but progress in decreasing worldwide poverty levels has slowed down due to the COVID-19 pandemic. In this paper, we analyzed the government of [redacted] attempt to help people out of poverty through the [redacted] cash transfer program. Previous research on other countries' programs suggest that traditional methods to measure poverty, such as household surveys, are expensive and time consuming. Therefore, we extend state-of-the art attempts to predict poverty in [redacted] via an existing framework proposed by Marty and Duhaut (2021).

Specifically, we perform a two-stage modeling process where stage 1 is composed of a transfer learning approach using 5 VGG16 pre-trained CNN deep learning models to predict nighttime levels with multi-spectral satellite imagery. In stage 2, we extract the learned features from the CNN models and merge them with the household survey data. This data was used as inputs in the extensive grid-search of supervised machine learning models to predict binary and continuous asset-based poverty measures.

After implementing our changes, we achieved a 3.8% overall validated accuracy improvement (44.9% to 48.7%) over the two-stage base model in predicting the binary asset-based poverty measure. This improvement in accuracy translates to predicting about 128 more households correctly in terms of poverty or no poverty. Along with accuracy, we also saw improvements in precision. Additionally, we modified the framework to accommodate a continuous poverty measure (as opposed to only a binary poverty measure). Although the regression performance metrics were not ideal (R^2 of .0319), we do not have a comparison model since the original model did not implement a continuous poverty model.
Overall, our changes to the model predicting a binary poverty outcome measure improved poverty prediction. We can attribute these improvements to the addition of multi-spectral bands via multiple CNN models as well as expanding the grid-search over a broader range of hyperparameters. Overall, our findings confirm that this is still a challenging problem to solve even with advanced modeling techniques, but our improvements are encouraging to help motivate future research.

There are a few limitations or caveats to our findings. Our CNN models did not have significant improvements in performance over the pre-trained models. This is a consequence of several factors. First, the pre-trained model architectures are large and were trained on over millions of images. Due to keeping nighttime light bin classes balanced, we needed to sub-sample from the majority NTL bin classes which removed many images from Landsat and VIIRS. Our largest input data to the CNNs only contained approximately 40,000 satellite images. This is not enough input data to make significant improvements to the pre-trained models. Secondly, even if we could use hundreds of thousands of images to train deeper CNN models than the pre-trained VGG16 model, it is extremely computationally expensive

to train all 5 CNN models on large set of images. Additionally, it should be noted that we decreased the recall score which effectively decreased the F-1 score. A decrease in recall score implies that we increased the amount of false negatives. Therefore, the updated model incorrectly classifies impoverished households to be financially stable at higher rates than the original model. It is up to the discretion of the World Bank and the government of [redacted] to decide which performance metrics are most important to the success of the [redacted] program

Given our work, future research should continue to use data science approaches in predicting poverty levels, as opposed to the strictly traditional methods discussed earlier. For researchers with access to high computing power, we recommend to use a similar two-stage modeling framework. Preferably, one that would gather hundreds of thousands of daytime satellite imagery and input not just RGB bands, but other spectral bands such as infrared or ultraviolet light into several CNN models to predict levels of nighttime light. Then for each observational unit where poverty is measured (household or neighborhoods), use many different data sources (household surveys, Facebook, mobile data) along with the features learned from the CNN models to predict a known poverty score via machine learning models in the 2nd stage. This approach currently is the standard for attempting to accurately measure poverty and future research can expand upon this by implementing similar techniques on additional, richer data sources.

Attribution Statement:
Laniah contributed to the expansion of the binary continuous grid search, the creation and analysis of the asset indices, and analysis of individual assets as potential proxies for poverty. Taymour researched different methods of incorporating multispectral imagery into the model, edited code for prepping the CNN and [redacted] data, and prepped CNN and [redacted] data for the CNN model. Michael prepped the CNN and [redacted] data, researched using pretrained CNN models for single band images, and modified the CNN feature extraction portion of the model to work with single channel models. Rus added 4 additional convolutional neural network models for single channel bands, edited the code to merge all extracted features from the 5 CNN models with the survey data, edited the PCA code to run on the larger merged cnn, survey data frame, and constructed the README file.

References

Aizenman, N. (2017). How to fix poverty: Why not just give people money? *National Public Radio (NPR)*.

Baird, S., C. McIntosh, and B. Ãzler. National health sciences research committee in malawi (protocol number: 569) and by the human research protections program committee at university of california.

Banerjee, A., P. Niehaus, and T. Suri (2019, 8). Universal basic income in the developing world. *Annual Review of Economics 11*, 959–983.

Burke, M., A. Driscoll, D. B. Lobell, and S. Ermon (2021, 3). Using satellite imagery to understand and promote sustainable development.

Fatehkia, M., B. Coles, F. Ofli, and I. Weber (2020). The relative value of facebook advertising data for poverty mapping. pp. 934–938. AAAI press.

Gertler, P. J. and S. Boyce (2001, 4). An experiment in incentive-based welfare: The impact of progresa on health in mexico.

Ghosh, T., S. J. Anderson, C. D. Elvidge, and P. C. Sutton (2013). Using nighttime satellite imagery as a proxy measure of human well-being. *Sustainability (Switzerland) 5*, 4988–5019.

Helber, P., B. Bischke, A. Dengel, and D. Borth (2017, 08). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012). Measuring economic growth from outer space.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016, 8). Combining satellite imagery and machine learning to predict poverty. *ScienceMag 353*, 790–794.

Kondmann, L. and X. X. Zhu (2020). Measuring changes in poverty with deep learning and satellite images.

Marty, R. and A. Duhaut (2021). Predicting poverty from the sky.

Steele, J. E., P. R. SundsÃžy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. EngÃž-Monsen, Y. A. D. Montjoye, A. M. Iqbal, K. N. Hadiuzzaman, X. Lu, E. Wetter, A. J. Tatem, and L. Bengtsson (2017, 2). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface 14*.

The World Bank (2020). World bank poverty report. Technical report, International Bank for Reconstruction and Development / The World Bank.

Tingzon, I., A. Orden, K. T. Go, S. Sy, V. Sekara, I. Weber, M. Fatehkia, M. GarcÃ-a-Herranz, and D. Kim (2019, 12). Mapping poverty in the philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. Volume 42, pp. 425–431. International Society for Photogrammetry and Remote Sensing.

Vyas, S. and L. Kumaranayake (2006, 10). Constructing socio-economic status indices: how to use principal components analysis. *Health Policy and Planning 21*(6), 459–468.

Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke (2020, 12). Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications 11*.

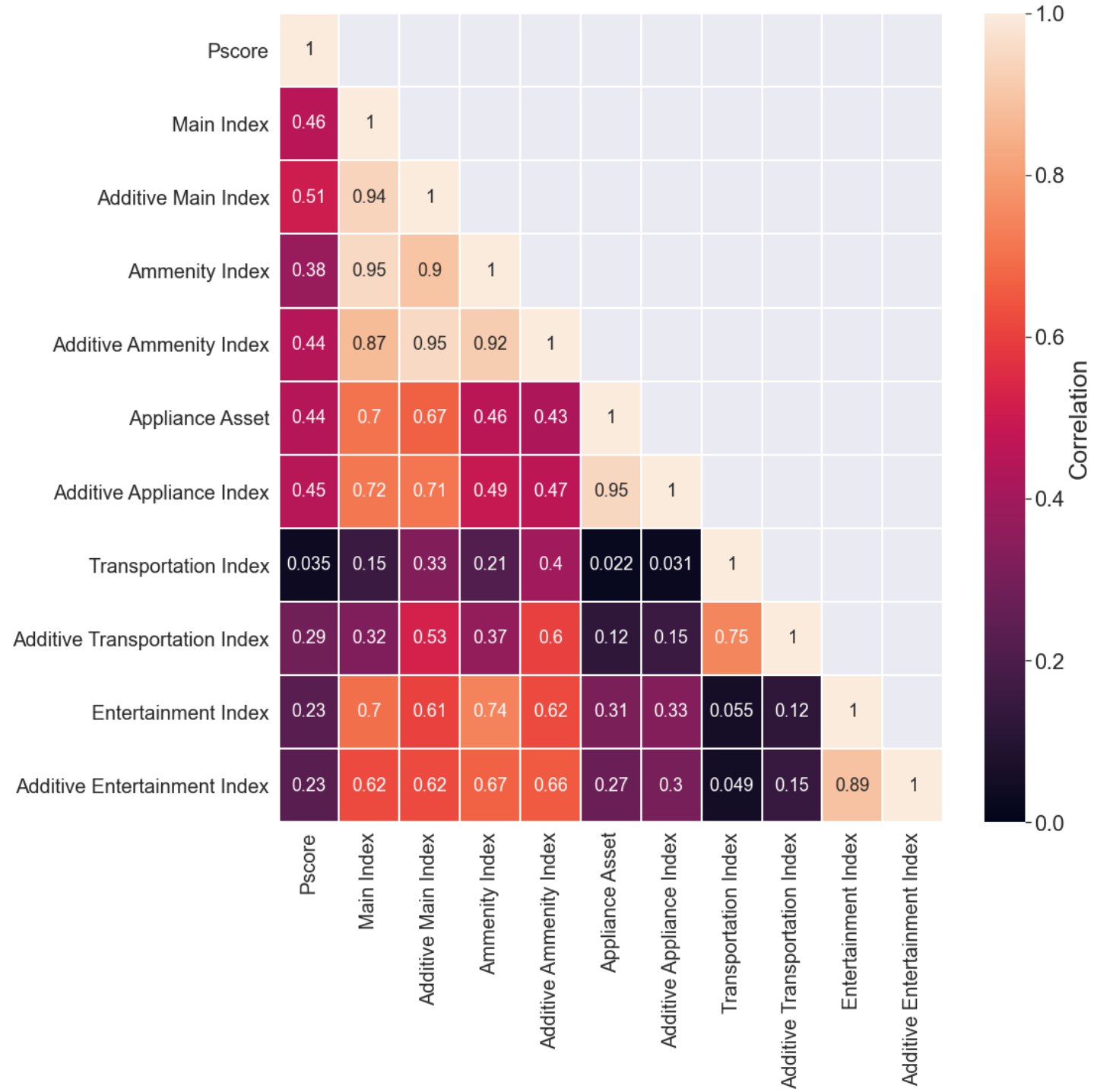

Figure 1: This chart displays the correlation between different asset indices as well as the poverty score