Temporal Poverty Prediction in Developing Countries

Ruchir Rastogi — rrastogi@stanford.edu

I. Introduction

Poverty reduction is the first of seventeen Sustainable Development Goals that the United Nations hopes to achieve by 2030 [1]. A prerequisite to reducing poverty levels though is understanding how wealth levels are currently distributed throughout a region and how they might change in the future. Using both current information and forecasts, governments and NGOs can then effectively allocate their limited resources.

Unfortunately, in many developing countries in Africa, poverty data is particularly scarce: countries conduct surveys infrequently and collect data from only a limited portion of their population [2]. The goal of this project is to reduce this data gap using publicly available data sources. In particular, we train convolutional neural networks (CNNs) to estimate poverty levels using satellite imagery, as aerial footage of a region can identify major landmarks of wealth (e.g. crop fields and buildings).

We train our deep learning model to not only estimate poverty levels in a specific year but to also predict changes across years. Each type of prediction serves its own purpose. Single year predictions help fill in data for regions where no surveys were conducted. Temporal predictions can be used to forecast how the distribution of poverty will change over time. Both ultimately would be instrumental in helping policy makers improve their country's macroeconomical decisions.

II. Previous Work

There has been a flurry of recent work that has attempted to use deep learning techniques to analyze satellite imagery. Newly published papers have estimated a wide range of economic and political factors, such as urban land use and terrorist financing strategies, that are oftentimes too hard to quantify directly [3] [4]. The success of these efforts has shown that it is possible to extract good features from high-resolution, aerial imagery.

Our current work is most closely related to previous work done in the Ermon lab that similarly attempted to predict poverty levels in Africa [5]. This current work differs from that in several crucial ways. First, Jean *et al.*'s original paper used high resolution (5 meters/pixel) Google Static Maps imagery, but since Google's imagery is proprietary, we have utilized much lower resolution imagery (30 meters/pixel) from the Landsat 7 satellite that is publicly available. In addition, the original paper focused on making predictions in a single year, whereas we focus on making predictions over time. Lastly, we also use a different source of poverty data—Living Standard Measurement Study (LSMS) surveys—since only LSMS surveys visit the same households every year. This is crucial for garnering any temporal insight.

However, from the original paper, we do adopt the important insight of using a transfer learning approach. Since LSMS surveys are very small in scope (they only collect data from a few thousand households each year), we cannot train a neural network model directly on poverty data. Instead, we first train a model to predict nightlight intensity (i.e. a region's brightness when visualized by a satellite at night), which is a proxy of wealth. Then, we take information from the neural network and feed it into a much less complex model to generate a final poverty score. This transfer learning approach is described in more detail below.

III. Data Collection and Preprocessing

A. LSMS Poverty Data

Poverty data was sourced from surveys conducted by the Living Standards Measurement Study (LSMS), which is made publicly available by The World Bank [6]. These surveys provide longitudinal panel data describing the wealth and consumption levels of approximately 3000 households in each country.[1] Both wealth and consumption levels are economic indicators of poverty. Wealth describes the total value of the assets a household owns, while consumption details their annual expenditures.

Current experimentation has focused on four developing countries in Africa: Malawi, Nigeria, Tanzania, and Uganda. These countries were chosen because

[1]The World Bank approximates wealth and consumption levels using qualitative survey questions.

surveys were conducted in each in at least three different years. The original data was averaged at the latitude/longitude level so that we could appropriately extract the location's corresponding satellite imagery. This preprocessing step reduced the number of data points per country to approximately 500.

B. Landsat 7 Satellite Imagery

Publicly available satellite imagery was taken from the United States government's Landsat 7 satellite program [7]. Unlike proprietary image sources such as Planet or Google Static Maps, Landsat 7 imagery is low resolution with each pixel covering a 30m $\times$ 30m region. Landsat 7 imagery does, however, extend for a long period of time (since 1999) and contain hyperspectral bands, such as infrared and thermal bands, not provided by other data sources.

A number of preprocessing steps were performed on the raw data. First, the government captures pictures monthly, but since LSMS poverty data is collected on a yearly basis, we create annual composites by taking the median value across all months. Secondly, the RGB bands were pan-sharpened so that they have a 15m $\times$ 15m resolution. Lastly, up to 5% of the pixels in a given year's imagery are NANs (not a number). We get rid of NANs by setting them equal to the band's mean value.

C. VIIRS Nightlights Data

Nightlight intensities were taken from the National Oceanic and Atmospheric Administration's Visible Infrared Imaging Radiometer Suite (VIIRS) [8]. The VIIRS sensor records daily information about an area's brightness at night at a spatial resolution of 750 meters. Recordings are scaled and normalized to the range $[0, 63]$. The data is already averaged per year, so no further preprocessing was required.

IV. Methods

A. Single Year Model

We use deep learning to predict wealth and consumption levels in the aforementioned four countries. However, since the amount of data points for which we know ground truth values (approximately 500 per country) is so few, we use a transfer learning approach where a CNN is first trained to predict nightlight intensity (discretized into three bins corresponding to low, medium and high) from daytime Landsat 7 imagery. Nightlights serve as as a good proxy for wealth levels because the two are generally strongly correlated. Intuitively, one would expect that downtown San Jose appears brighter at night than a farm in rural Alabama. The last layer of the CNN, which encodes a feature representation of the original image, is then fed into either a ridge regression or gradient boosted trees (GBT) model in order to predict wealth levels. Figure 1 describes the model used to estimate poverty levels in a specific year.

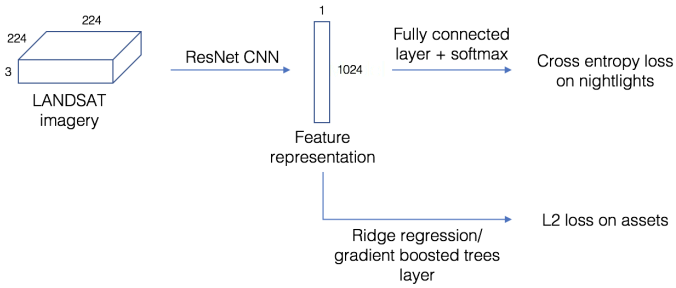


Fig. 1: Model structure for single year estimates

B. Temporal Model

For our temporal model, we use a very similar transfer learning approach. However, instead of training on values from a specific year, we train the CNN to predict the change in nightlights (again binned into three categories corresponding to low, medium, and high change) using the difference in satellite imagery across two years. For all four countries, we try to predict changes in wealth between 2009 and 2013. Figure 2 illustrates the temporal model.

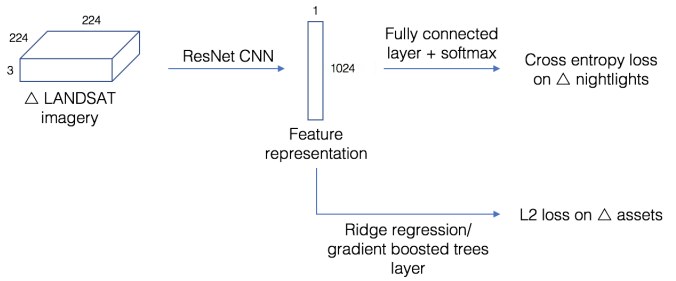


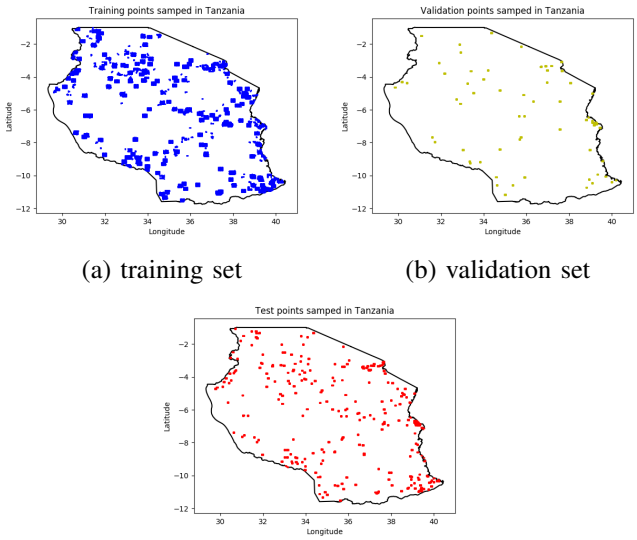
Fig. 2: Model structure for temporal estimates

C. Data Sampling

As described earlier, we know ground truth wealth and consumption levels for very few data points: roughly 500 per country. Since the LSMS surveys are longitudinal studies and collect information from the same households every year, this same data shortage exists for the temporal model as well. To circumvent this problem, we first set aside all 2000 locations where

the LSMS survey was administered for the small linear regression/GBT model at the end.

In order to train the CNN on nightlights data, we sample 1km $\times$ 1km regions broadly from the four countries, making sure not to overlap with any of the LSMS locations. However, we cannot sample locations randomly because doing so would lead to an unbalanced dataset. Recall that in the single year model, nightlights were binned into three categories (low, medium, and high). Similarly, in the temporal model, the change in nightlights was binned into the same three categories. Since approximately 70% of locations would be placed in the low category for either the single year or temporal model, we selectively sample locations that would be placed in the medium and high bins in order to achieve a balanced dataset. We also samples images more densely near areas for which we have ground truth poverty labels, as that has been shown to lead to better results [5]. This approach yields roughly 100 thousand data points that we split 70-10-20 into our training, validation, and test sets. Figure 3 illustrates the data split in Tanzania:



(a) training set (b) validation set

(c) test set

Fig. 3: Data split in Tanzania

D. Convolutional Neural Network

We use a residual network, as popularized by He *et al.*, that contains shortcut connections between layers [9]. The network is pretrained on the ImageNet dataset, as studies have shown that the general objection detection capabilities of pretrained networks help speed up learning for satellite imagery analysis [5]. All neural network code was written in TensorFlow.

E. Ridge Regression and GBT

The last portion of our model involves feeding the final layer's activation into a ridge regression or gradient boosted trees (GBT) model. Models of low complexity are used because we only have 2000 LSMS data points to train them on.

We use a ridge regression model that contains L_2 regularization instead of a naive linear regression model to prevent us from overfitting on our small data set. In addition, gradient boosted trees, which in short are an ensemble of weak learners trained on reducing the errors of previously generated learners, are tested to see if introducing non-linearities into the model will help [10]. We use 5-fold nested cross validation to tune the regularization hyperparameter in the ridge regression model and the maximum depth and maximum features hyperparameters in the GBT model.

V. Results and Discussion

A. Metrics

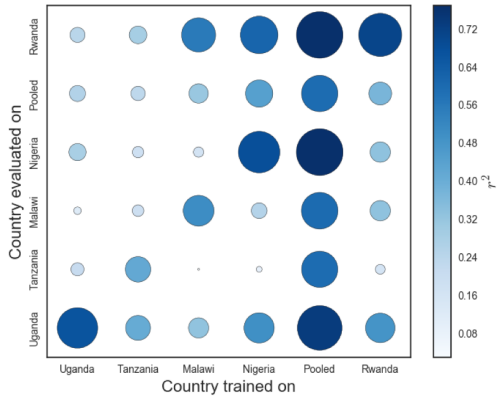
We test the performance of our models by calculating the r^2 value (coefficient of determination), which denotes the amount of variance in the underlying data that is captured by our predictions.

B. Single Year Model Optimizations

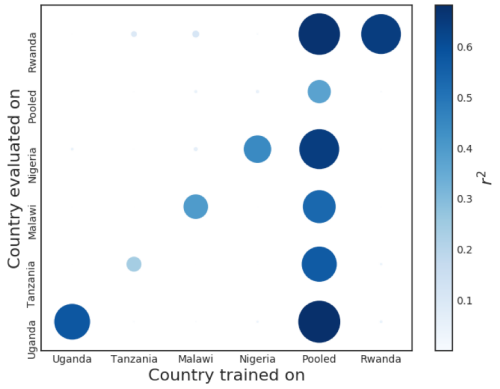
Since our model is trained on satellite imagery that contains non-RGB bands unlike the Google Static Maps imagery used in Jean *et al.*, we first determine whether these additional bands improve model performance. As shown in Figure 4, models trained on non-RGB bands do not generalize well to other countries. (In this experiment, the final ridge regression layer was trained on the higher-fidelity Demographic and Health Survey poverty data used in Jean *et al.*'s original work, which explains why the r^2 values are also correspondingly better than those reported in the rest of this paper.)

These results clearly indicate that the model tends to pick up on features in non-RGB bands that are specific to a particular country. Accordingly, for further experiments, we disregard non-RGB bands.

We then tuned the optimal depth of the ResNet by examining the model performance of an 18-layer network and a 50-layer network (both respectively optimal with regards to their regularization parameters). As the results in Table 1 illustrate, the 18-layer ResNet is significantly better. The additional layers in the



(a) RGB bands



(b) Non-RGB bands

Fig. 4: Mean cross-validated (n = 5) test r^2 values for predicting asset levels

deeper network likely pick up on features that are weakly correlated with nightlights but not with poverty levels. In addition, the 18-layer network outperforms our baseline, which is a ridge regression model that predicts wealth levels using a scalar nightlight intensity as input.

	Assets	Consumption
18-layer ResNet	0.47	0.31
50-layer ResNet	0.35	0.26
Nightlights (baseline)	0.38	0.19

TABLE I: Mean cross-validated (n = 5) test r^2 values

C. Temporal Results

Having optimized the single layer model, we now test whether our temporal model can predict the difference in wealth in a region across two years. Table II displays the results. Crucially, these results likely underestimate the true values because the temporal model's hyperparameters were taken from optimal single year model and are not tuned specifically for this task due to time constraints.

	Δ Assets	Δ Consumption
Model (linear)	0.19	0.12
Model (GBT)	0.25	0.16
Δ Nightlights (linear)	0.07	0.04
Δ Nightlights (GBT)	0.15	0.09

TABLE II: Mean cross-validated (n = 5) test r^2 values

Our temporal model comparatively outperforms our Δ nightlights baseline, but does not compare to our single year model. Introducing non-linearities through the use of gradient boosted trees did however significantly help.

These results are relatively surprising because the CNN's training and test set loss decreased normally during the training loop; accordingly, we do know that the model successfully detected changes in nightlight intensities given satellite imagery. However, our model's relatively poor performance can be attributed to the poor performance of the Δ nightlights baseline. The entire premise of the transfer learning approach rests on the assumption that the proxy we train the CNN on will allow the network to indirectly decipher features relevant to determining changes in poverty levels. But if changes in nightlights are themselves not correlated with changes in wealth, then this knowledge transfer cannot occur.

D. Error Analysis

To figure out if there were other reasons why our temporal model struggled to learn, we first attempted to determine which examples it made poor predictions on. As Figure 5 indicates, our temporal model does decently in low and high-income regions, but worse in those areas closer to the mean. Manual inspection of the data reveals that changes in wealth levels tend to be higher in places that are already very poor or very rich (unfortunately, in opposite directions as income inequality has spread). This suggests that the model is doing a good job extracting features that detect changes, but cannot learn features that indicate lack of much change.

In addition, Principal Component Analysis was used to determine how many of the 512 features extracted by the ResNet for each test image actually varied. Unfortunately, as Figure 6 shows, our data lies approximately on a 10-dimensional subspace, as only

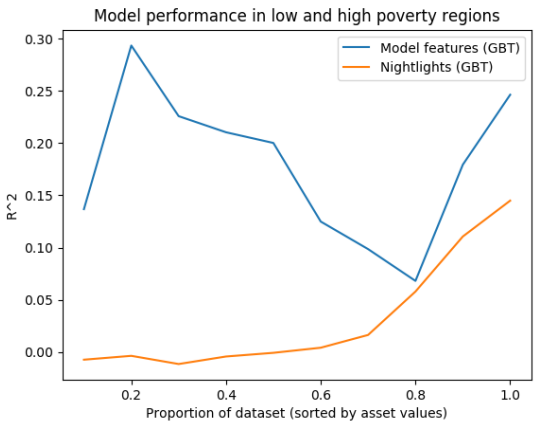


Fig. 5: Temporal model performs significantly better in low and high-income regions

10 principal components account for the vast majority of the variance. This indicates that our model is not picking up on good, separating features when being trained on changes in nightlights.

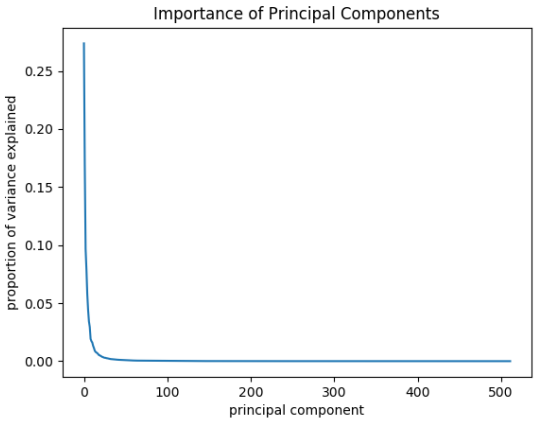
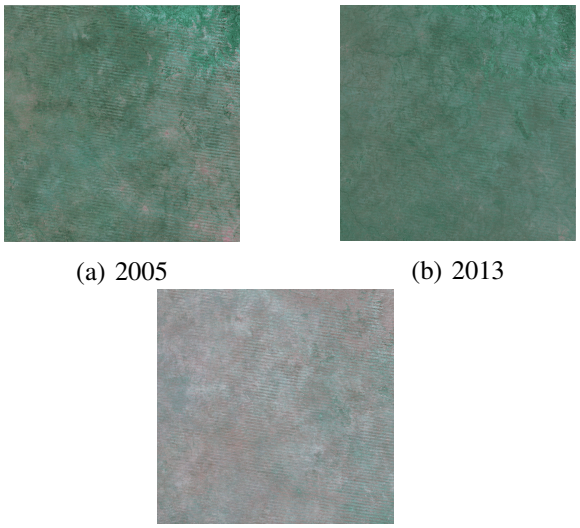


Fig. 6: Proportion of variance in data explained by each principal component

Lastly visual inspection of the low-resolution imagery reveals just how hard it is for the model to pick up on distinguishable features that change over time. Figure 7 shows two zoomed out pictures of Kampala, Uganda's capital city, in 2005 and 2013. There is practically no difference visible to the human eye. Accordingly, it is not surprising that our model does not perform nearly as well as one would expect.

VI. Conclusion and Future Work

In this work, we have shown the possibility of using low-resolution, publicly available satellite imagery to



(a) 2005 (b) 2013

(c) Difference

Fig. 7: Landsat Images of Kampala over time

predict poverty levels both in single years and across multiple years. While we can successfully predict poverty levels in specific years, there is significant work that remains to be done on the temporal side before it can help guide policymakers' decisions. Specifically, either we must find better proxies than nightlights, as changes in nightlights do not correlate with changes in wealth, or the entire transfer learning methodology must be scrapped. In addition, it is likely that better quality imagery will be needed before temporal models will work well.

Once we reach that point though, the applications of this work are endless. We can conceive of models in which a recurrent neural network (RNN) is placed on top of the CNN so that an arbitrary amount of previous years' satellite imagery can be used to make future forecasts. In addition, we could use semi-supervised models to encode the idea that poverty is spatially correlated to make our predictions more precise. Using a combination of these approaches could allow us to achieve our ultimate goal: accurately predicting poverty levels at a 30m $\times$ 30m resolution that extends both into to the future and into the past.

VII. Acknowledgements

I would like to thank the Sustainability and Artificial Intelligence Lab at Stanford for providing these datasets and for guiding me along the course of my project. I would like to specifically thank Professor Ermon for allowing me to work in his lab the past few months and for being more than gracious with his help and time. Lastly, I would like to thank George Azzari and Anthony Perez for writing most of the code that allowed me to easily manipulate the giant TIFF files that contained Landsat 7 satellite imagery and VIIRS nightlights data.

References

[1] United Nations. *Sustainable Development Goals: 17 Goals to Transform Our World*. Available at: http://www.un.org/sustainabledevelopment/sustainable-development-goals/

[2] Kathleen Beegle, Luc Christiaensen, Andrew Dabalen, and Isis Gaddis. *Poverty in a Rising Africa*. The World Bank. 2016.

[3] Adrian Albert, Jasleen Kaur, and Marta C. Gonzales. *Using Convolutional Networks and Satellite Imagery to Identify Patterns in Urban Environments at a Large Scale*. 2017. Available at: https://arxiv.org/pdf/1704.02965.pdf

[4] Quy-Toan Do *et al. How Much Oil is the Islamic State Group Producing? Evidence from Remote Sensing*. The World Bank. 2017. Available at: http://documents.worldbank.org/curated/en/239611509455488520/pdf/WPS8231.pdf

[5] Neal Jean *et al*. *Combining satellite imagery and machine learning to predict poverty*. Science, 353. 2016. 790794.

[6] The World Bank. Living Standard Measurement Study. Available at: www.worldbank.org/lsms

[7] United States Geological Survey. Landsat Data Access. Available at: https://landsat.usgs.gov/landsat-data-access

[8] National Oceanic and Atmospheric Administration. VIIRS Daily Mosaic. Available at: https://ngdc.noaa.gov/eog/viirs/download_ut_mos.html

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. IEEE. 2015. Available at: https://arxiv.org/abs/1512.03385

[10] Alexey Natekin and Alois Knoll. *Gradient Boosted Machines, A Tutorial*. Frontiers in Neuroscience, 7. 2013. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/pdf/fnbot-07-00021.pdf