Measuring Poverty via Remote Sensing in an Unsupervised Setting

Matthew Mauer[1], *Ryan Webb*[1]

[1]University of Chicago. Harris School of Public Policy.

mrmauer@uchicago.edu, ryanwebb@uchicago.com

Abstract

Poverty mapping in underdeveloped countries can be extremely difficult because their institutions do not have the capacity for regular intake and maintenance of survey and administrative data. Many existing solutions that are applied in development organizations rely on a transfer learning model using cheap and widely available satellite images first proposed by Jean et al[1] in 2016, while the most recent literature takes a reinforcement learning approach using extremely high-resolution satellite images that are more expensive, but can be used in combination with coarser resolutions to improve the learning process[2].

Alternately, we propose an unsupervised framework for poverty mapping using a Convolutional Variational Encoder-Decoder to encode a Landsat-8 image and decode it to a nighttime lights image. Features are extracted from the final layer and fed to a downstream regression or classification task. We find promising early results, and our model is competitive with the original transfer learning approaches of Jean et al by some metrics. We believe our model provides a proof of concept that supervised methods may be substituted for unsupervised ones with regards to poverty mapping. However, future work will be needed to further optimize the model and to learn more about the nature of the features being extracted.

Index Terms: development economics, unsupervised machine learning, computer vision, representation learning, geographic information systems(GIS)

1. Introduction

Measuring consumption levels in underdeveloped countries is done predominantly via statistical analyses on small scale survey and administrative data. The resulting measurements are considered to be the most accurate and trustworthy information on the topic. However, survey data is expensive to collect, and waves are only done sporadically. This process leaves an increased chance of decisions being made on out-of-date estimates. In recent years, models using remote sensing have been developed and are constantly improving. They often rely on satellite imagery, which is cost-effective, global, and able to provide new estimates in a nominal amount of time. For poverty mapping, both Landsat and nighttime lights satellite images are used in the learning process[1][3][2].

The amount of light visible by satellite at night has emerged as an important metric in the social sciences for measuring economic development[4]. Conveniently, it is an easily understood, common-sense indicator: we expect areas with higher economic development to have higher nighttime light radiance because it is reflective of goods like electricity, industry, and infrastructure, and the personal and societal wealth to have such things. The proliferation of satellite imagery allowed the creation datasets like ImageNet that offered huge sets of labelled Landsat images, as well as constantly updating worldwide nighttime lights images (VIIRS) collected by NOAA.[3] These datasets, in turn, led to the creation of a transfer learning approach to using satellite images to predict poverty[1].

Similar implementations of that work been developed and deployed by the World Bank[5]. These approaches use an ImageNet dataset to pre-train a Convolutional Neural Network (CNN) to directly extract socioeconomic factors for poverty classification, then then change the classifier on the last layer to a nighttime light classification task. They train again on the new task, using ImageNet parameters as initialization to achieve knowledge transfer[1]. In this iteration, a feature vector from the final layer is used as input to a logistic regression on poverty.

We propose that an unsupervised approaches to this problem could be competitive with this foundational transfer learning method. We believe this could be particularly helpful to the research community as a viable unsupervised approach to feature extraction would eliminate the need for pre-trained models. Specifically, we believe that an adaptation of a Convolutional VAE model is a natural fit for the spirit of this task. Remember, the fundamental economic concept at the center of this model is nighttime lights, and we believe that a Conv VAE that encodes an RBG-8 Landsat image then decodes it to an equivalent nighttime lights representation can be used to learn both the Landsat image and nighttime lights image, resulting in a simpler process for feature creation and extraction to the downstream task.

2. Data

We have three main sources of data: Landsat 8 RGB images for daytime images of the land area (Google Static Maps), Visible Infrared Imaging Radiometer Suite[4] (VIIRS) for nighttime images of the land area with light radiance (NOAA), and Living Standards Measurement Study (LSMS) survey data from the World Bank. Each respondent in LSMS is associated with a geographic cluster of $10km^2$, for which we retrieve 1 VIIRS image of dimenstionality $21x21$ and a Landsat 8 image from around the centroid collected at dimensionality $400x400$. Our survey data is from the year 2016 while our Landsat and VIIRS images are from the years 2016 and 2015 respectively. Our data universe consists of LSMS respondents from the countries of Malawi, Nigeria, and Ethiopia.

2.1. Remote Sensing

2.1.1. Daytime Landsat

Landsat images are images covering a pre-defined land area. In most cases, and ours, composite Landsat images are used that are Landsat images collected from privately-held satellites and freely available using the Google Static Maps API. We use the three RGB bands from these images as inputs to our model. We collect these images and transform them to matrices, where they are resized from dimensions $400x400$ to $256x256$. We coarsen the data to make it computationally efficient and learnable with the much lower dimension VIIRS data.

It is important to note here that, while it is theoretically possible to use high-resolution Landsat data from a provider like Sentinel-2, as in Ayush et al, we do not have the computational resources necessary to process such data at scale nor the reinforcement learning framework that makes such analysis cost-effective.

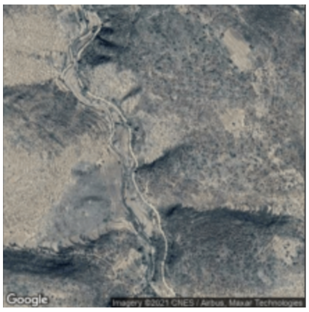

Figure 1: *Example image from our Landsat collection*

2.1.2. *Visible Infrared Imaging Radiometer Suite*

VIIRS data captures a measure of radiance during the nighttime horus at the pixel level, where pixels comprise roughly $1.7km^2$ of space[4]. As with our Landsat data, we are working with a composite image over the period of six months to ensure that no pixels are subject to cloud cover that obfuscates the level of radiance. This data is freely available from NOAA, and was acquired as a tif file of the world, broken up into tiles, which were read in as needed depending on the country of interest. In preprocessing, specific sub areas of each tile are mapped to their respective clusters of LSMS respondents[6].

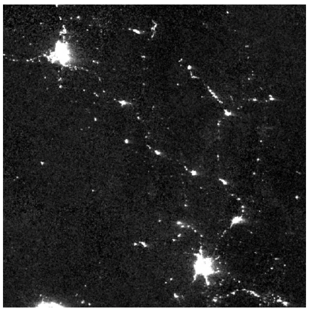

Figure 2: *Example image from our VIIRS collection*

2.2. Living Standards Measurement Survey

The LSMS is administered by the World Bank, and collects data on many dimensions of household and individual well-being to assess household welfare, understand household behavior and evaluate the effect of various government policies on the living conditions of people in low- and middle-income countries[6]. We collect the data for our relevant countries from the World Bank's public datasets. We derive the measures of aggregate consumption using the same methodology as Jean et al (2016) and encode an indicator variable for poverty based on a \$1.90 (USD) per day consumption threshold. This variable is our predicted value in downstream regression and classification tasks.

3. Model

Our work seeks to improve upon the representation learning methods of previous approaches while duplicating the structure of the downstream classification task for comparability of representation learning methods. Our representation learning model adapts the Variational Autoencoder (VAE) proposed by Kingma and Welling (2013) to use both the Landsat daytime satellite imagery and VIIRS nighttime light data.

3.1. Representation Learning - Convolutional VED

Instead of the traditional VAE that maps an input to a latent space and the latent space representation to a recontruction of the input (sometimes with denoising or other augmentation), our model encodes the data rich daytime imagery to a low-dimension latent space and maps the latent space representation to a reconstruction of the nighttime lights image of the same geographic region. Because of this discrepancy with traditional VAE, we choose to instead call our model a Variation Encoder Decoder (VED). It is variational because we apply the same reparamaterization trick in the latent space suggested by Kingma and Welling[7]: The final hidden layer of the encoder maps to an equal number of means and variances conditional on the input $\mathbf{x}$

$$\mathbf{z} \sim \mathcal{N}(\mu, \sigma|\mathbf{x}) \tag{1}$$

that are then used to sample from before decoding to an estimation of the corresponding VIIRS image. Also taken from Kingma and Welling is the ELBO loss function slightly modified and used for training our VED. The ELBO loss function here combines the KL divergence of $q(\mathbf{z}|\mathbf{x})$ and the $\mathcal{N}(0, 1)$ while using the second view (i.e. the VIIRS image) to calculate the reconstruction loss.

$$L(\mathbf{x}, \mathbf{z}, \mathbf{y}) = -D_{KL}(q(\mathbf{z}|\mathbf{x})||\mathcal{N}(0,1)) + E_{q(\mathbf{z}|\mathbf{x})}(log(p(\mathbf{y}|\mathbf{z}))) \tag{2}$$

For the reconstruction loss term, we use Binary Cross Entropy, where the pixel average radiance values in $\mathbf{y}$ have been normalized to $(0, 1)$ by a natural logarithm transformation and divided by the natural logarithm of the maximum radiance value in the data set (i.e. $ln(94,000) = 11.45$).

We also use convolutional and transpose convolutional layers to better encode and decode spatial components of both the Landsat and VIIRS imagery. Specifically, we use four convolutional layers in the encoder before passing to two fully connected linear layers. From the latent representation, the decoder begins with a single hidden layer and passes along to three transpose convolutional layers. The lower complexity of the decoder relative to the encoder is to account for the lower resolution of the VIIRS image used.

We use a Leaky Rectified Linear Unit (Leaky ReLu) non-linear activation function at all convolutional layers. We chose the Leaky ReLu to speed up training and to avoid the dying-neuron problem where a node ceases learning due to a persistent zero activation value. For the output layer, we use the sigmoid activation function as our VIIRS pixels are normalized to $(0, 1)$ and we use Binary Cross Entropy for our reconstruction loss.

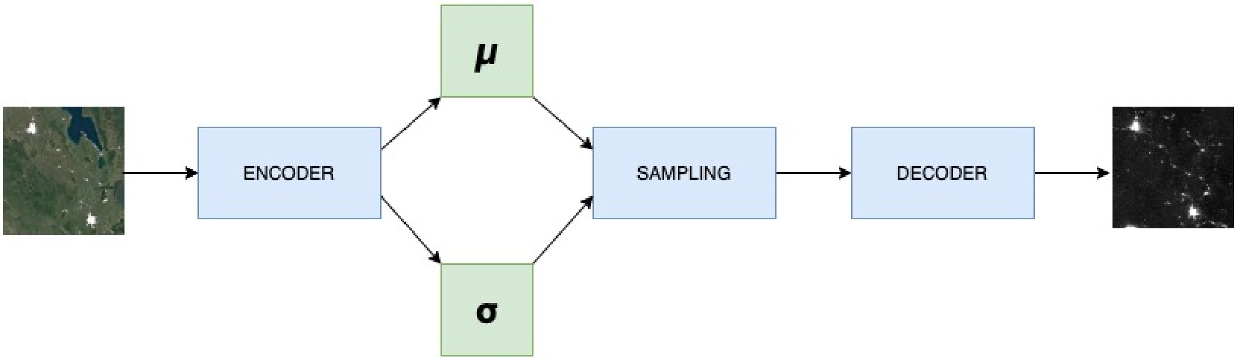

Figure 3: *Diagram of VED.*

3.2. Predicting Poverty

For all downstream tasks, the representation used is the $\mu|\mathbf{x}$ that we obtain by passing a Landsat RGB image into the encoder. The representation vectors obtained from our trained Convlutional VED (ConvVED) are then used in two separate but similar down stream tasks.

The first downstream task is a regression on the learned representation to predict a continuous cluster-average daily household consumption. While this task hasn't been performed in the literature assessing GIS features for predicting poverty, we deemed it useful to measure the variance in consumption captured by our learned representation which we can estimate with the R^2 of our overall model's performance on the test data. The specific model used for this task was a linear Elastic Net. By tuning the hyperparameters of our Elastic Net, we can test for varying degrees of model sparsity. This is helpful in understanding the number of learned features that are in fact useful for the prediction task.

More in line with previous research, we also use the learned features for a binary classification of poverty by Logistic Regression. Similarly to our regression task, we use an Elastic Net Variant of Logistic Regression, so that we may tune for model sparsity by cross-validation.

3.3. Learning

For learning the feature representation, all variants of hyperparameter combinations were trained with 10 epochs on the entire training set (81% of full set) and early stopping was performed using a validation set (9% of the full set). We chose to use Adam optimization for learning in hopes that the momentum helped our model converge more quickly to global minimum given the limited size of our data.

We tune for optimal hyperparameters including the weights placed on the KL Divergence regularizer, the weight on the variance within the KL Divergence regularizer, the number of components in the latent space, and the learning rate. Optimal hyperparameters are determined by the latent representation's downstream performance in the supervised model for consumption and poverty prediction. Our hyperparameter search included a wide range of possible dimension for the latent space including 32, 64, 128, and 256 dimension iterations. The optimal values are discussed in the Results section.

Both our linear and our logistic supervised models were trained and tuned for optimal L1 and L2 weights using SAGA[8], an optimization method for solving composite objective functions.

4. Results

4.1. Poverty Prediction

The results of downstream poverty classification for the optimal representation model are shown in Table 1. We see that our ConvVED model outperforms raw predictions by nighttime lights and Xie et al's representation on the F1 score and is competitive with Xie et al's approach measuring by AUC. This comes with the caveat that the probability threshold chosen for poverty classification (0.2) in our model was chosen by tuning the threshold on the F1 score. Nonetheless, it is hopeful to see the ConvVED compete Xie et al's transfer learning representation given the discrepancy in both model complexity and training data volume. Xie et al's was trained using a data set of 370,000 images compared to the less than 25,000 in our set. There's also the question of model generality. Xie et al trained and tested their model using image from only Uganda, while ours incorporated a wider variety of topographies from West, East, and Southern Africa (Nigeria, Ethiopia, and Malawi respectively). Xie et al also used higher resolution images with a frame about one tenth the size. This means their CNN may have been able to capture more granular feature that ours was not.

	Survey	**Lights**	**Transfer**	**ConvVED**
F1	0.552	0.448	0.489	**0.507**
Accuracy	0.754	0.526	0.716	0.579
Recall	0.722	0.914	0.658	0.873
Precision	0.450	0.298	0.394	0.357
AUC	0.776	0.719	0.761	0.742

Table 1: *ConvVED are the results obtained on the Convolutional VED model, while Survey are the results using data obtained from manual survey, Lights are the results using only the nighttime light intesity for poverty predictions, and Transfer are the results from Xie et al transfer representation learning approach. All non-ConvVED figures come from Xie et al.*

Results from the linear regression predictions of household consumption cannot be directly compared to previous work but are illuminating for how much information the learned features capture. The R^2 for the optimal model measured on the held out test set was $\mathbf{0.12}$. While this immediate result seems to indicate that our feature representation may not be a strong measure of variation in consumption, it is possible that they could play an useful role as an input to a larger stacked model integrating several data sources.

4.2. Optimal Hyperparameters and Model Density

The optimal hyperparameters for our feature representation included 64 dimensions in the latent space and KL divergence and variance regularizers with a weight of 1.

Interestingly, tuning of the KL divergence and variance regularizers did not have consistent effects on the downstream model performance and was generally dependent on the size of the latent space dimensionality. If a trend from our grid search results can be discerned, it is that the model benefited from a higher KL Divergence regularizer and a lower variance regularizer.

In tuning the L1 loss ratio for both the regression and classification supervised models, we found that sparse models were favored: L1 ratios ranged from 0.9 to 1. This was especially the case for our regression model where in some iterations only a small fraction of the features were used for predicting consumption. This seems to indicate that the ConvVED learns several features that, while they produce nighttime radiance, are not indicative of the local levels of consumption. This could include pass through infrastructure such as lighted roads or rail lines that pass nearby an impoverished village but do not contribute to local production and industrial activity. Other alternative explanation include water bodies or other reflective surfaces that shine moonlight toward the VIIRS satellite. This last point emphasizes that improved pre-processing steps such as filtering out moon reflection may yield better results.

5. Conclusions and Future Work

We see our work as having delivered a proof of concept that unsupervised methods are applicable to the poverty prediction task at hand. Our model performs reasonably well with minimal data preprocessing and optimization and considers only one unsupervised method. We see future work as iterating on two parallel fronts, data and model selection, with aspirational goals that will be briefly addressed in our final statements.

Our work looks only at our predictions, without delving into what indicators are recovered with feature extraction or relative model performance on different land areas. As such, immediate next steps should be learning more about these features and further analyzing our results. There is some basic work that can be done, including creating activation maps, that can help give us intuition about what features our model is learning. That information would enable us to more thoughtfully tune our model's hyperparameters and also to learn more about where we might want to use additional data to compliment our Landsat/VIIRS images.

For example, recent work from Ayush et al[2] takes advantage of using small samples of high-resolution Landsat images to improve performance of the transfer model proposed by Jean et al[1] on urban areas, where the model traditionally performs poorly. We expect that our model would similarly perform less well on urban areas and would benefit from a similar form of data augmentation.

We are also interested in exploring other unsupervised learning methods. We have naturally multi-view data, and we are specifically interested in implementing a Variational CCA model[9], where we could learn more explicitly the relationship between a daytime image and a nighttime image by framing each band as a view for the model.

All recommendations of future work up to this point are achievable in immediate to near-term, but in the medium term these models could benefit from having a time series component for the model to learn from. Our model collects images in time t and classifies for time t, but if remote sensing poverty models could be used to classify for time $t + 1$ they have the potential to become part of the general purpose modeling discussion for poverty in underdeveloped countries.

6. Acknowledgements

We would like to than Karen Livescu and Freda Shi for their guidance during the course of our research. Also instrumental was code used for preprocessing GIS data that was borrowed and only slightly modified from Jatin Mather[10] at the University of Illinois.

7. References

[1] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016. [Online]. Available: https://science.sciencemag.org/content/353/6301/790

[2] K. Ayush, B. Uzkent, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Efficient poverty mapping using deep reinforcement learning," 2021.

[3] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping," *arXiv e-prints*, p. arXiv:1510.00098, Sep. 2015.

[4] M. Z. C. D. Elvidge, K. E. Baugh and F.-C. Hsu, "Why viirs data are superior to dmsp for mapping nighttime lights."

[5] M. Rama and R. Beyer, "Measuring south asia's economy from outer space," 2017.

[6] W. B. Group, "Living standards measurement study (lsms) brochure," 2020. [Online]. Available: http://documents1.worldbank.org/curated/en/708961597206589588/pdf/Living-Standards-Measurement-Study.pdf

[7] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.

[8] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," 2014.

[9] W. Wang, X. Yan, H. Lee, and K. Livescu, "Deep variational canonical correlation analysis," 2017.

[10] J. Mather, "Predicting poverty replication," https://github.com/jmather625/predicting-poverty-replication, 2020.