Poverty Estimation Using Satellite Images and Geospatial Data

Christine Cepelak
christinecepelak@gmail.com

Janine De Vera
janinepdevera@gmail.com

Johannes Halkenhäußer
halkenjo@gmail.com

Abstract

This project[1] *aims to contribute to the research body on bridging poverty statistics measurement gaps, through the application of machine learning techniques in satellite imagery and open source spatial data. The model building experiments conducted thus far focus on constructing and operationalizing a data pipeline that will allow poverty prediction through the use of daytime satellite images. The main dataset was constructed from combined geotagged wealth indices and satellite images for one country (Malawi). This was then used to run a simple baseline linear regression and random forest models. While preliminary results currently reveal a lot of room for improvement in terms of explanatory power, additional data and further fine-tuning of the models are expected to improve performance. The project will be expanded to include more data from three additional countries, a convoultional neural network, as well as refined model specifications.*

1. Proposed Method

The scope of the study includes four (4) different countries: Ethiopia, Malawi, Mali, and Nigeria. For purposes of the midterm report, only the methods and results of Malawi are presented in detail.

The methodology used for preliminary tests and experiments is discussed in two parts - (i) data preparation and (ii) model building.

1.1. Data preparation

The main dataset used in this study was constructed by merging satellite image data and geospatial wealth indices. Prior to combining the datasets, significant pre-processing had to be applied to the satellite data in order to remove duplicate and overlapping images.

Satellite data often comes in a set of thousands of images, with each image a tensor of square pixels, and each pixel a three or four-dimensional array. The images for Malawi are composed of 256 x 256 pixels with RGBA (Red, Green, Blue, and Alpha for transparency) features. Each image therefore had 262,144 (256 x 256 x 4) different attributes.

Principal Component Analysis (PCA) (trained on a random sub-sample) is applied to reduce dimensionality of the data into 20 components. This smaller dataset is then filtered for image duplicates. The image indices identified through the PCA-reduced data are combined with the geotagged wealth indices to create the final merged dataset, which is then divided into training (85%) and test (15%) datasets.

1.2. Model building

After construction of the training dataset, a PCA is again used to project in a lower dimensional space while at the same time maximizing the variance of the input data [4]. This PCA is separately trained from the PCA used for filtering. Its output is used to train further models.

Linear Regression: A linear regression model minimises the Mean Squared Error (MSE) by finding an optimal combination of coefficients that are used to build a linear combination of the input features. The linear regression is not adjusted in any way, giving an indication of how the most simple model fares.

Random Forest: 51 Random Forest Regressors (RFR) with 4 folds each (204 fits) are trained using the merged Malawi data. A RFR grows individual Regression Trees given some parameter inputs (described in 2.3.C) that minimises the MSE using binary recursion and splitting the data along multiple decision nodes. The individual trees are ”pruned” (i.e. optimised in size to prevent overfitting) and combined into forests. Generally, larger forests allow for more robust estimators. By layering a 4-fold cross validation on top of each parameter combination, overfitting is again tested against [2] [4]. Enabled by the scikit-learn function for grid searching with cross validation, it was not necessary to create validation sets manually.

[1] GitHub respository: https://github.com/ccepelak/ML-SS22

2. Experiments

2.1. Data

A. *Ground truth data (wealth index)*

The main outcome variable for the study was collected from the Demographic and Health Surveys (DHS) Program. The DHS is regularly conducted in over 90 countries, with the purpose of gathering representative data on health and economic well-being of different populations.[1, 2, 4, 5]

The wealth index as reported in the DHS is calculated as the first principal component of several household attributes which are not direct measures of economic status. Examples of variables that are accounted for in the index are sources of drinking water, types of bathroom facility, ownership of various household appliances (e.g. refrigerator, television, telephone, etc.), and materials used for building the house. The indices are calculated based on pre-defined clusters that comprise several households. Cluster centroids are reported as the mean latitude and longitude of the households that belong to one cluster group.

Figure 1 shows the distribution of wealth indices across the 850 clusters of Malawi. This will later on be compared to the predicted wealth indices of the different models.

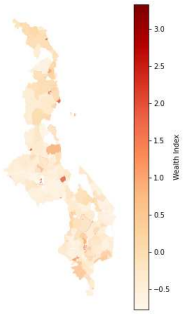

Figure 1. Malawi ground truth wealth index

B. *Satellite images*

The satellite images dataset was created from Planet Developer Resource Center's Planet and retrieved from Kaggle API[3]. Each image is made up of 256 x 256 pixels, with RGBA values. These images are matched to wealth indices based on the location of cluster centroids as defined in 2.1.A.

The spatial matching of wealth indices and images yielded both n:1 and 1:n correspondence. That is, there

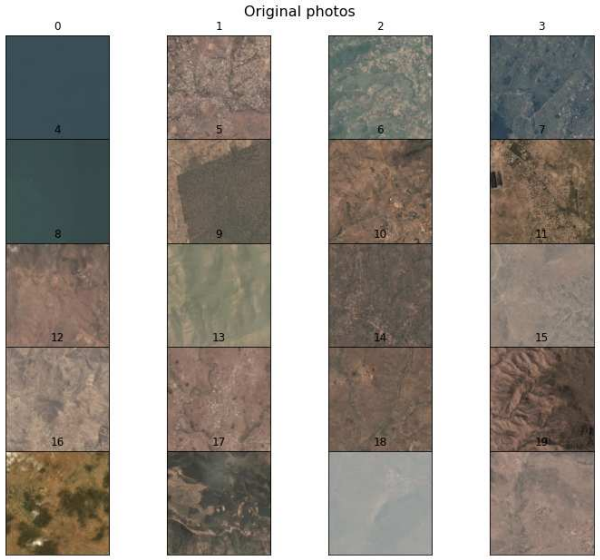

Figure 2. A sample of satellite photos of Malawi

are cases where (i) several wealth indices are matched to one image and (ii) several images are matched to one wealth index. Matching yielding a training set of 13,337 matches and 451 unique images with 10,568 unique outcomes.

Figure 3 shows the overlap between the outcome variable and satellite image data for Malawi. The grey clusters pertain to areas which have representative satellite images, while blue clusters are those that have have representative wealth indices. The distribution of images across the country is unequal, with more images available in regions with more survey results, which are likely densely populated areas.

Figure 3. Overlap of satellite images and DHS wealth indices

Table 1 describes the main datasets used in this study. Information for the other three countries are also presented.

Demograpic and Health Survey				
	Ethiopia	*Malawi*	*Mali*	*Nigeria*
Year collected	2019	2016	2018	2018
No. of clusters	305	850	379	1,400
No. of households	8,663	26,361	9,510	40,427
Daytime Satellite Imagery				
	Ethiopia	*Malawi*	*Mali*	*Nigeria*
Year collected	2015	2015	2015	2015
No. of images	8,587	12,700	12,800	11,535

Table 1. Summary of wealth index and satellite image datasets.

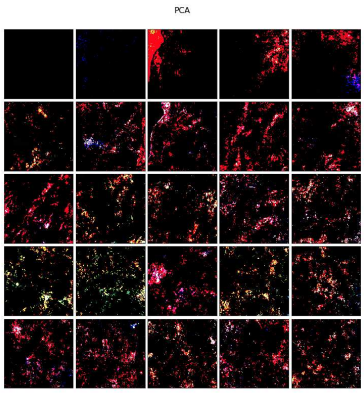

Figure 4. The representation of each of the 25 components produced by the PCA. They explain roughly 90% of variance within predicted value.

2.2. Evaluation Method

As in related studies, the R^2 is used an evaluation metric to compare the performance of different models. The R^2 measures the variance in the dependent variable (wealth index) that is explained by the independent variables. For purposes of the midterm report, only the main explanatory variable of satellite images is used to predict poverty.

2.3. Experimental Details

A. *PCA*

Given the large size of the image data, principal component analysis (PCA) was used to reduce the size of the images down to 25 components. The PCA used whitening which de-correlates the outputs. This PCA is different to the PCA that was trained for filtering the data, so that only training data is used. Given the computational constraints of local machines a sub-sample of 1,000 random images was used to train the PCA (a full sample will be used when the entire pipeline is moved to the server). After this point, all experiments are being run with the same PCA reduced training data set. Figure 4 shows a representation of the 25 components. It highlights how the models have to pick up on patterns counter intuitive to the human eye. Not assembles of villages/water etc. are identified but patterns of colors.

B. *Linear Regression*

As the most simplistic baseline model we ran a linear regression with the PCA-reduced full training set and no additional parameters.

C. *Random Forest*

To account for the non-linearity of the data, a non-parametric RFR was used. All forests use MSE as the splitting criterion and use bootstrapped samples. To test out multiple hyper-parameters, cross validated with 4 folds grid search went over all different combinations of following parameters:
max_depth:
The maximum tree depth determines how many levels the tree is allowed to have. The deeper the tree is allowed to grow the larger the risk overfitting, as a deeper tree also leads to a higher segmentation of the outputs. *Range:* [1,2,3,4,5, 10, 16, 18, 20, 22, 24, 26, 30, 40, 50, 100, 200]
max_features:
During the training, the model can take into account a varying number of features. This furthers decorrelation of the trees and thus allows to build a more robust forest. During the grid-search, allowing the full number of features, the square root of the number of features, and the log_2 of the number of features were passed to for the model parameters. With 25 samples, the $log_2(25) = 4.64$ is roughly $\sqrt{25} = 5$, so no difference can be expected. However, the distinction is kept in for future use should a higher number of PCA components be employed. *Range*: [n_features, $\log_2$(n_features), $\sqrt{n\_features}$]

2.4. Results

The linear regression has performed poorly and in the latest test only had a R^2 of 5.64% on the training set. The result is very low and with a higher performance on the test set, its lack of robustness to new data is evident.

The results for the grid search of various models of the RFR are shown in Figure 5. The best estimator is chosen and reported for overview in Table 2.

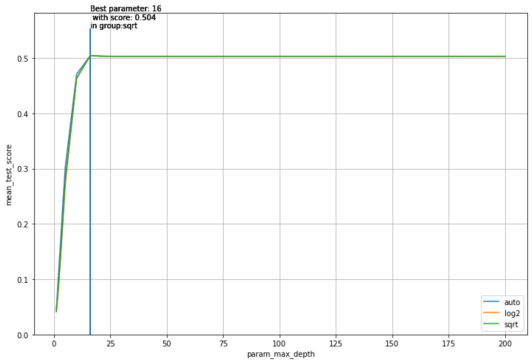

Figure 5. R^2 score for different parameter combinations, averaged over the CV.

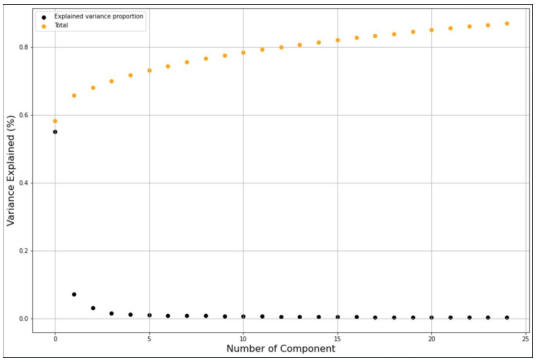

Figure 6. The marginal variance explained decreases for each component added.

Method	Training R^2	Test R^2
Literature [4]	-	70%
Linear Regression	5.64%	7.02%
Random Forest	50.4%	49.1%

Table 2. Test and training score for each method. For random forest the best estimator is reported with a maximum tree depth of 16 and the square root of features being considered.

2.5. Discussion

The results currently produced by our models are not on the level of the results achieved by the papers in the project proposal.

The linear regression results reveal the shortcomings of pure linear methods in the complicated non-linear environment of computer vision. That the test score is higher than the training score gives testament to the inability of the model to explain the data effectively. It was to be expected that a linear regression would perform quite poorly but still this result is disappointing.

The Random Forest has performed much better than the linear model because its non-parametric nature allows it to deal with the non-linearity. The cross validation when testing forest parameters, bootstrapping of individual trees, and defining a maximum number of features to be considered allowed to find a robust forest that did not overfit. The standard deviation of the CV-scores for the best estimator was 0.7%. Hence, the test score lies within 2 standard deviations of the training-score of the best parameter.

Regardless of the specific parameters given to the forest, increasing the number of trees will make it more robust and improve the scores. The following general improvements are necessary:

1. *PCA-dependency:*
   The results are highly dependent on the quality and effectiveness of the PCA. Currently, the PCA is trained only with a random sub-sample of the data. Increasing the number of samples to train the PCA (ideally, all) and also the number of components will improve its ability to represent the data. Increasing the sample size should increase the effectiveness of each component and increasing the number of components will allow the models to train on more dimensions. The marginal variance explained decreases with each additional component, so adding adding additional becomes insignificant after some time (Figure 6).

2. *More data:*
   The current models are only being run with Malawi country data, hence only on a sub-set of our data. This choice was made as the computational resources required for image processing make the process slow and lengthy if all images and countries are included. The focus is currently on establishing a full pipeline before ramping up the sample size. While increasing the sample size will make the models more robust, adding more countries adds further variation, requiring further training for the models.

3. Future work

Future work will include:

- Application of the pre-processing pipeline to all countries and constructing a larger dataset;
- Possible inclusion of other geospatial covariates to improve explanatory power of models;
- Construction of a Convolutional Neural Network (CNN);
- Application of linear regression, random forest, and, ultimately, CNN to the full data set; and
- Continued hyper-parameter tuning.

For efficient application of the models, we will also explore some methods (based on feedback from Hertie faculty, Prof. Lowe) in dealing with the high number of features in the full dataset: Kernel Trick, Random Projection, Iterative Reweighted Least Squares (IRLS).

References

[1] X. Chen and W. D. Nordhaus. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594, 2011.

[2] C. Ledesma, O. L. Garonita, L. J. Flores, I. Tingzon, and D. Dalisay. Interpretable poverty mapping using social media data, satellite images, and geospatial information. *arXiv preprint arXiv:2011.13563*, 2020.

[3] san__bt. Satellite images to predict poverty, Jan 2021.

[4] X. Zhao, B. Yu, Y. Liu, Z. Chen, Q. Li, C. Wang, and J. Wu. Estimation of poverty using random forest regression with multi-source data: A case study in bangladesh. *Remote Sensing*, 11(4):375, 2019.

[5] Z. Zhongming, L. Linong, Y. Xiaona, Z. Wangqiang, L. Wei, et al. Mapping poverty through data integration and artificial intelligence: A special supplement of the key indicators for asia and the pacific. 2020.

Index of comments