2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) | 978-1-6654-0382-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ECTI-CON51831.2021.9454768

Machine Learning Approach with Multiple Open-source Data for Mapping and Prediction of Poverty in Myanmar

Nyan Lin Htet* Waree Kongprawechnon* Suttipong Thajchayapong†

Tsuyoshi Isshiki‡

***Abstract*—Poverty is rampant and very crucial issue in developing countries. Therefore, in this paper, we explore the implementation of machine learning on the estimation of poverty by training input data from widely available and accessible open-source, including nighttime lights (NTL) and OpenStreetMap (OSM) data. We propose this approach as a straightforward, cost-effective and alternative option for previous studies which have been done by deep learning. We applied the linear regression and ridge regression algorithm as our baseline models while using random forest regression, gradient boosting regression and xgboost regression to achieve the better performance. We found that our best model can explain approximately 74% of the variation in wealth index from input features of Myanmar. We then created the poverty map in province administrative level for Myanmar, which indicates that conventional machine learning models with open-source data can still be as efficient as deep learning on poverty estimation.**

***Index Terms*—Wealth index, poverty map, nighttime lights, openstreetmap, machine learning, Myanmar.**

I. Introduction

Sustainable Development Goals (SDGs) of the United Nations for the global in 2030 aim at eliminating poverty. It is essential to have reliable data on poverty to evaluate the lives of the vulnerable and devise poverty mitigation policies, both by decision makers and scholars efficiently. However, developing countries such as Myanmar where poverty is one-fifth of population [1], data scarcity is the major challenge for such tasks. Above all, poverty measurements traditionally build on ground surveys while collecting surveys are time-consuming, labor intensive, and costly. Subsequently, such methods can be done only every 3 to 5 years [2] and only one survey was carried out in 57 countries to generate poverty statistics between 2002 and 2011 [3].

In order to fill these gaps, remote sensing data have become publicly available by offering global-scale and been widely applied with machine learning by many researchers. Nighttime lights (NTL) data and high-resolution satellite images are two keys remote sensing data in poverty assessment. Among them, a more popular approach has been nighttime lights (NTL) data, and there are two common sources to acquire NTL data; (1) Defense Meteorological Satellite Program's Operational Line Scan System (DMSP-OLS) and (2) the Suomi National Polar-orbiting Partnership Visible Infrared Imaging Radiometer Suite (VIIRS) [4]. Elvidge et al. [5] established the global poverty map from DMSP data whereas Yu et al. [6] predicted poverty at the country level in China by NPP-VIIRS data. In recent years, researchers have tried to extend the functionality of nighttime lights data by incorporating machine learning. Jean et al. [7] proposed a 8-layer convolutional neural network model (VGG-F) with high-resolution daytime satellite images and nightlight data to approximate economic livelihood indicators.

Since then, there have been many papers which extend the method of the research [7] with different deep learning architectures such as GoogLeNet [8], ResNet-18 [9], ResNet-50 [10] and DenseNet [10]. Moreover, there are also some studies that utilized another types of non-traditional data such as call detail records (CDRs) [11], [12], crowd-sourced geographic information from OpenStreetMap (OSM) [14] and social media advertising data [15] in mapping socioeconomic indicators.

In this study, we propose an alternative methodology for deep learning approach from previous studies [7] - [10] because deep learning is costly computation, laborious and requires a large amount of data which is challenging for data scarce developing countries. Although most of the papers did not report any details of computational criteria, Pandey et al. [16] mentioned that his deep learning model to predict poverty, was trained for 192 hours (8 days) on an NVIDIA TITAN X GPU. Furthermore, accessing high-resolution RGB imagery is USD 10-20 per km^2 [17] and so it would cost a lot to cover the whole country. In order to save time and cost, we applied only freely available open-source data utilizing on machine learning models. To the best of our knowledge, this is the first study that explores the estimation of poverty in Myanmar using multiple open-source data on machine learning.

This paper is organized in the following way. Section II is about the details of our proposed methodology which is followed by our experimental results in Section III. Discussion

* School of Information, Computer and Communication Technology (ICT), Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand; m6222040351@g.siit.tu.ac.th

† National Electronics and Computer Technology Center 112 Thailand Science Park, Phahon Yothin Rd., Klong I, Klong Luang, Pathum Thani 12120, Thailand.

‡ Tokyo Institute of Technology, Ookayama Campus 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8550, Japan.

978-0-7381-1127-8/21/$31.00 ©2021 IEEE

is described in Section IV and conclusion is in Section V.

II. Proposed Methodology

The overall methodological framework of the study to solve the problem is illustrated in Fig. 1. Firstly, we obtained open-source data; DHS, NTL and OSM data as shown in the Table I. We then extracted input features which are independent variables, from NTL and OSM data by the geographical locations from DHS survey, to estimate wealth index; the lower values are, the higher level of the poverty is. The actual wealth index is provided by DHS survey which became supervised data while training the machine learning models upon the input features. We trained on different models because it is impossible to know beforehand which algorithm has the best performance. To evaluate the performance of our models, we utilized mean squared error (MSE) and coefficient of determination (R^2). Finally, as an output, a poverty map was generated from the model that has the best R^2.

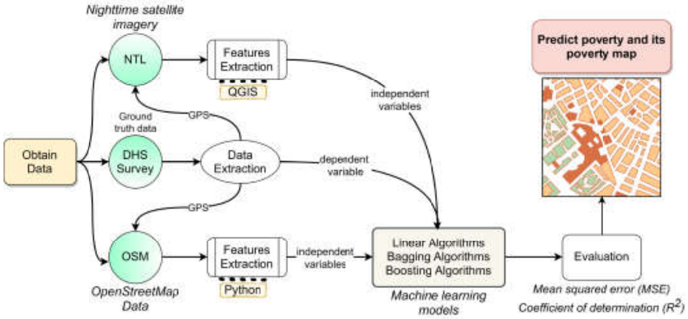

Fig. 1. Overall flowchart of methodological framework.

TABLE I
Primary Datasets of the Study

Source of data	Year collected	Coverage
Demographic & health survey (DHS)	2015-2016	12500 samples 441 clusters
Nighttime lights (NTL) data	2013 & 2016	the whole country
OpenStreetMap (OSM) data	2018	the whole country

A. Features Extraction

1) DHS data: In DHS data, there is positional error added by DHS on the geographical locations of the survey respondents in clustering formats (20-40 household samples in one cluster), which is around 5km and 2km for rural and urban areas accordingly, to cover their identities and privacies [18]. Hence, there are 441 clusters from DHS and we added noise to GPS coordinates by creating buffer zones for each cluster as a means to extract features more accurately from open-source data.

2) NTL features: In this study, we intended to use nighttime lights directly to indicate socioeconomic status in Myanmar because it is technically a good indicator for welfare as higher level of luminescence, are correlated with economic activity levels, investment on infrastructure, prosperity and electricity usage [19]. For this reason, we extracted luminosity levels of nighttime lights from both NPP-VIIRS and DMSP-OLS in maximum, minimum, mean, median, standard deviation for each cluster.

3) OSM features: In order to achieve the better performance of the model, nighttime lights data alone are not adequate due to their noisy nature [10], with less capability of distinguishing differences in wealth of very poor people [7]. Therefore, new features such as roads, buildings, land-use and point-of-interest (POIs) are extracted from OSM data because infrastructures and land-use are advocated with sustainable growth of the regions which enable us to partially determine the wealth of the regions [20]. For each cluster buffer, we extracted road features (types of roads; length of each kind of road; distance from the center of the cluster to the nearest road etc.), buildings (types of buildings, building area, mean distance from clusters), land use (types of land use, area size, proportion of area to buffer), and point-of-interest (count of each type of point-of-interests). As a result, we have about 288 features in total from our multiple open-source data.

B. Machine Learning Models

After deriving necessary features, we developed machine learning models to estimate wealth index and poverty areas. The purpose of deploying machine learning models is to predict the continuous target variables t given the value of a D-dimensional vector x of input variables [21]. In our machine learning models, there are 3 parts; linear algorithms, bagging algorithms and boosting algorithms. Among them, linear algorithms are baseline model since they are simple and readily applicable algorithms with low computational power [21]. We chose linear regression and ridge regression from linear algorithms. Mathematically, linear models for multiple input features can be expressed as

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + ... + w[i] * x[i] + b \qquad (1)$$

where $x[0]$ to $x[i]$ is the features in which i is the number of features while $\hat{y}$ is the prediction the model, and w and b are parameters of the model that are learned. Since linear models are easy to overfit due to training on many input features [21], we utilized bagging and boosting algorithms to enhance the performance. Therefore, we selected random forest regression, gradient boosting and XGboost regression. To incorporate these models, we used the popular scikit-learning library that offers a solid API for the creation and evaluation of machine learning models.

C. Model Performance Evaluation

To evaluate the regression models, we employed the coefficient of determination (R^2) metric which provides the goodness of fit of a set of predictions to the actual values, and its value is between 0 and 1 for no-fit and perfect fit on the regression line. It can also be interpreted as how much percentages of outcome variance is explained by the input variables in the model. R^2 is mathematically defined as

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

where $\bar{y}$ shows the mean of all y_i, $\hat{y}_i$ is the estimated value of the i-th test sample, and y_i is the ground-truth value for total n samples. Another metric to check the loss of the model is mean squared error (MSE) that can measure the average squares of the errors between the estimated values and the actual values. The value closer to zeros indicates better performance, can simply be described as

$$\text{MSE}(y,\hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \tag{3}$$

where y_i refers to the ground-truth value, and $\hat{y}_i$ represents the predicted value of the i-th test sample.

III. Experiment and Result

Initially, we standardized all input features to improve the efficiency of models by scaling numerical input variables to a standard range. Then, we checked and compared the performance on estimation of wealth index between DMSP-OLS and NPP-VIIRS. We split the data into 67:33 ratio for training and testing respectively. The 10-fold cross validation from Jean et al. [7] was also adopted. Although NPP-VIIRS data are more reliable in other study [4], we found that DMSP-OLS provides better result in our case study as demonstrated in Fig. 2. For this reason, we chose DMSP-OLS NTL data to merge with OSM data. Nighttime lights data alone can reveal just almost 55%-68% of the variance in the target variable. Finally, both NTL and OSM features were trained upon the machine learning models and the best performance is achieved by the gradient boosting regression as displayed in Fig. 3, resulting in the 74% of explicable variance in wealth index of Myanmar. This demonstrated that the combination of features from multiple data sources can explain the poverty better than the features from only one source of data.

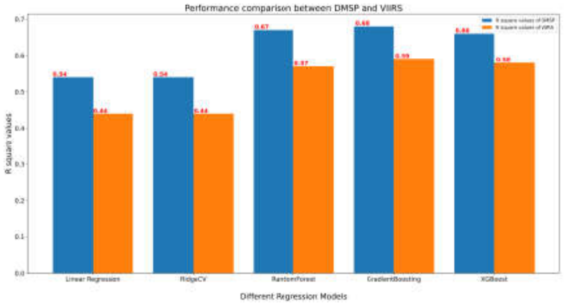

Fig. 2. Models performance from DMSP-OLS and NPP-VIIRS.

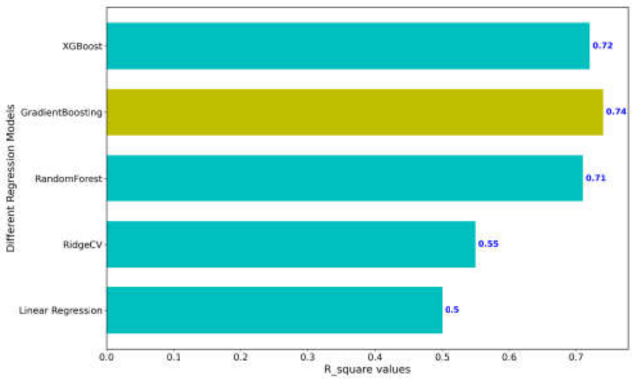

Fig. 3. Models performance from merging OSM and NTL features.

Furthermore, we also reported the MSE for all constructed models as shown in Fig. 4 in order to confirm that our selected model has the least error rate too. In this study, the lowest MSE score was achieved by the gradient boosting regression model, which is 0.15 and its loss decreased along the training and testing iterations as displayed in Fig. 5, showing the learning ability of the model. We fine-tuned and optimized the parameters and hyperparameters by grid search method as shown in the Table II to achieve the best performance. Therefore, our fine-tuned selected final model, the gradient boosting regression model, has the lowest error rate with the highest performance of R^2 value.

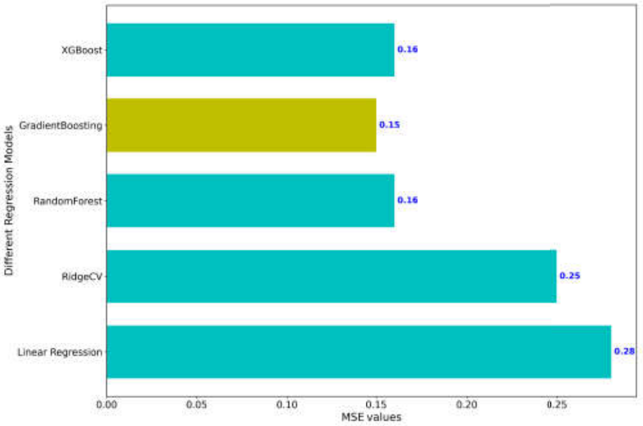

Fig. 4. Mean square error (MSE) for each model.

Visualization of the predicted and the ground-truth wealth index in Myanmar is displayed in Fig. 6, and a poverty map in province level is generated for Myanmar as shown in Fig. 7.

IV. Discussion

This study explores the application of open-source data and machine learning approach for poverty estimation without applying computationally complex and costly deep learning models. Our final model achieved the encouraging result of R^2, 0.74 in prediction on asset-based wealth index. Although

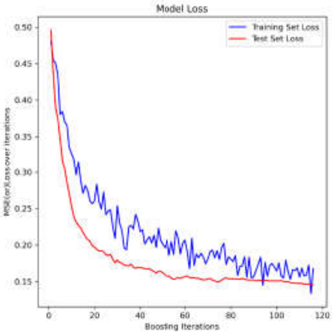

Fig. 5. Loss function of the model.

TABLE II
FINE-TUNED VALUES OF MODEL PARAMETERS

Parameter Name	Value
learning_rate	0.143
n_estimators	116
subsample	0.73
min_samples_split	6
min_samples_leaf	2
max_depth	1
random_state	40

we cannot yet conclude that our proposed method outperforms the previous studies [4], [7], [8], [14], [23] conducted in different countries, our final result is promising one in terms of R^2 comparing to these studies. Furthermore, we have only a few parameters to train without any cost for acquiring data which makes our approach easily be applicable and explorable for other countries. Additionally, it took only a few hours to extract features from open-source data and train the model on the ordinary graphic processor unlike the previous study which took days to train even with the expensive high performance processor [16].

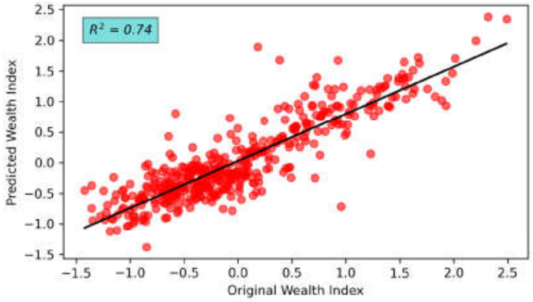

Fig. 6. R^2 between estimated wealth index and ground-truth wealth index in Myanmar

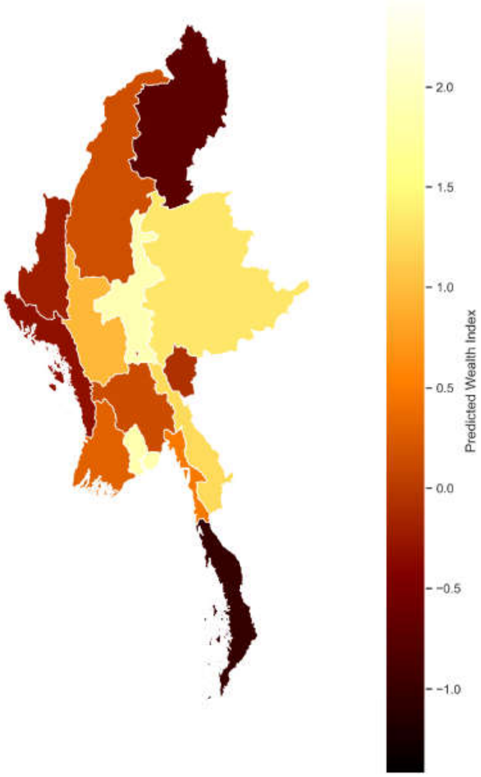

Fig. 7. Province-level poverty map for Myanmar

V. CONCLUSION

In this paper, we studied the combination of multiple open-source data by applying directly on machine learning models to estimate poverty. We demonstrated the usage of 5 different machine learning models and selected the gradient boosting regression as a final model in the case of Myanmar. In our work, the results showed that the machine learning models can explain the poverty better when the input features from multiple data were integrated together rather than using from one source of data. In other words, poverty is a complex phenomenon that cannot be measured accurately by one single data type. Moreover, the experimental results indicated that our proposed method can be potentially applicable to other countries, and it is also the reliable, time-saving, cost-efficient, and simply-usable option for poverty estimation in data scarce developing countries like Myanmar.

ACKNOWLEDGMENT

This research is financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology, Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU) under the TAIST Tokyo Tech Program.

References

[1] "Poverty Report-Myanmar Living Conditions Survey 2017," World Bank.https://www.worldbank.org/en/country/myanmar/publication/poverty-report-myanmar-living-conditions-survey-2017 (accessed Sep. 12, 2020).

[2] J. Klugman, Ed., A sourcebook for poverty reduction strategies. Washington, D.C: World Bank, 2002.

[3] U. Serajuddin, H. Uematsu, C. Wieser, N. Yoshida, and A. Dabalen, Data Deprivation: Another Deprivation to End. The World Bank, 2015.

[4] S. Piaggesi et al., "Predicting City Poverty Using Satellite Imagery," 2019, pp. 90–96, Accessed: Jul. 22, 2020. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/Piaggesi_Predicting_City_Poverty_Using_Satellite_Imagery_CVPRW_2019_paper.html.

[5] C. D. Elvidge et al., "A global poverty map derived from satellite data," Computers & Geosciences, vol. 35, no. 8, pp. 1652–1660, Aug. 2009, doi: 10.1016/j.cageo.2009.01.009.

[6] B. Yu, K. Shi, Y. Hu, C. Huang, Z. Chen, and J. Wu, "Poverty Evaluation Using NPP-VIIRS Nighttime Light Composite Data at the County Level in China," IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, pp. 1–13, 2015, doi: 10.1109/JSTARS.2015.2399416.

[7] 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, Aug. 2016, doi: 10.1126/science.aaf7894.

[8] B. Babenko, J. Hersh, D. Newhouse, A. Ramakrishnan, and T. Swartz, "Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images, With an Application in Mexico," arXiv:1711.06323 [cs, stat], Nov. 2017, Accessed: Jun. 20, 2020. [Online]. Available: http://arxiv.org/abs/1711.06323.

[9] C. Yeh et al., "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa," Nat Commun, vol. 11, no. 1, p. 2583, Dec. 2020, doi: 10.1038/s41467-020-16185-w.

[10] Y. Ni, X. Li, Y. Ye, Y. Li, C. Li, and D. Chu, "An Investigation on Deep Learning Approaches to Combining Nighttime and Daytime Satellite Imagery for Poverty Prediction," IEEE Geosci. Remote Sensing Lett., pp. 1–5, 2020, doi: 10.1109/LGRS.2020.3006019.

[11] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," Science, vol. 350, no. 6264, pp. 1073–1076, Nov. 2015, doi: 10.1126/science.aac4420.

[12] J. E. Steele et al., "Mapping poverty using mobile phone and satellite data," J. R. Soc. Interface., vol. 14, no. 127, p. 20160690, Feb. 2017, doi: 10.1098/rsif.2016.0690.

[13] M. von Mörner, "Application of Call Detail Records - Chances and Obstacles," Transportation Research Procedia, vol. 25, pp. 2233–2241, 2017, doi: 10.1016/j.trpro.2017.05.429.

[14] L. Zhao and P. Kusumaputri, "OpenStreetMap Road Network Analysis for Poverty Mapping," p. 7.

[15] M. Fatehkia et al., "Mapping socioeconomic indicators using social media advertising data," EPJ Data Sci., vol. 9, no. 1, p. 22, Dec. 2020, doi: 10.1140/epjds/s13688-020-00235-w.

[16] S. Pandey, T. Agarwal, and N. C. Krishnan, "Multi-Task Deep Learning for Predicting Poverty From Satellite Images," 2018.

[17] K. Ayush, B. Uzkent, M. Burke, D. Lobell, and S. Ermon, "Efficient Poverty Mapping using Deep Reinforcement Learning," arXiv:2006.04224 [cs], Jun. 2020, Accessed: Jun. 16, 2020. [Online]. Available: http://arxiv.org/abs/2006.04224.

[18] C. R. Burgert, J. Colston, T. Roy, and B. Zachary, "Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys," 2013, doi: 10.13140/RG.2.1.4887.6563.

[19] A. Bruederle and R. Hodler, "Nighttime lights as a proxy for human development at the local level," PLoS ONE, vol. 13, no. 9, p. e0202231, Sep. 2018, doi: 10.1371/journal.pone.0202231.

[20] Reshaping economic geography. Washington, D.C: World Bank, 2009.

[21] C. M. Bishop, Pattern recognition and machine learning. New York: Springer, 2006.

[22] K. Baugh, F.-C. Hsu, C. D. Elvidge, and M. Zhizhin, "Nighttime Lights Compositing Using the VIIRS Day-Night Band: Preliminary Results," APAN Proceedings, vol. 35, no. 0, p. 70, Jun. 2013, doi: 10.7125/APAN.35.8.

[23] I. Tingzon et al., "Mapping poverty in the Philippines using machine learning, satellite imagery, and crowd-sourced geospatial information," Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., vol. XLII-4/W19, pp. 425–431, Dec. 2019.