Coal Pit Mapping with Random Forest-Based Ensemble Machine Learning at Lower Benue Trough

Okeke Francis Ifeanyi[*], Ibrahim Adesina Adekunle[**], Echeonwu Emmanuel Chinyere[***]

[*] Department of Geoinformatics and Surveying, University of Nigeria, Enugu
[**] Department of Geoinformatics and Surveying, University of Nigeria, Enugu
[***]AMMI, African Institute for Mathematical Sciences, Rwanda

DOI: 10.29322/IJSRP.10.12.2020.p10851
http://dx.doi.org/10.29322/IJSRP.10.12.2020.p10851

Abstract- This work entails identifying areas of all coal pits in and around a part of Lower Benue trough where coal exploitation had taken place in the past, using random forest machine learning technique with out-of-bag (OOB) estimate of generalization error. Training sites were selected from the Landsat image covering the area of interest using QGIS image processing software. Two well-known exploited sites were selected as coal training sites while few other ones also identified were reserved for validation. Some other areas known to be different than coal sites were selected and classed as non-coal areas. the two sets of data were trained for binary classification using random forest ensemble supervised classification. Scikit-learn library run by python API was adopted for the machine learning. Open-source library, GDAL was employed to treat image raster and shape file within the python API. Accuracy of result was estimated to be equal to approximately 99.957%. A map showing location of all exploited sites was produced. Validation was carried out with known reserved coal pits observed during ground truth. The result accurately identified the validation sites and other areas properly.

Index Terms- *Coal pit, Machine Learning, Random Forest, 'OOB', Scikit-learn*

I. Introduction

Coal exploitation activity was known to be common within and around Benue trough. Official exploitation commenced in 1909 in Nigeria (Famuboni, 1996; Minjng, 2006) after discovery of coal, in the course of search for crude oil. Locally, activities of artisans were also common and many people did take advantage of abandonment of further formal exploration and exploitation for the resource after oil was struck and attention of the government was diverted to use of oil for powering railway activities and power generation, the then two major users of the product. The need for alternative source for powering electricity turbines as gas supplies dwindle and recent attention of the government to resuscitate the state's moribund iron and steel rolling company at Ajaokuta with attendant need for coal as raw material re-awaken the need to source for more coal. This work embraces inventory taking of coal exploitation sites hitherto opened for coal activities through mapping using ensemble machine learning technique.

II. Background

Machine learning is regarded as a field of study which engages implementation of algorithms in computer to transform data into intelligent action (Abellera et al., 2018).

Ensemble learning is a machine learning which aggregates several models to produce an estimator with performance metrics better than each single component. They include but not limited to Bagging (also called Bootstrap Aggregating), Bayes Optical Classifier, Boosting, Bayesian Model Averaging, Bucket of Models, Stacking and Random Forest.

Random forest combines several decision trees. A decision tree is a non-parametric supervised learning which is applied to both classification and regression (Albon, 2016). These trees are generated to strengthen the most popular class. Hence, a random forest classifier consists of "*a collection of tree-structured classifiers {f(x,Θ$_k$), k=1, ...} where the {Θ k } are iid random vectors and each tree gives a vote for the most likely class at input x"* (Breimen, L, 2001). With the ensemble of classifiers $f_1(x)$, $f_2(x)$, $f_3(x)$….$f_k(x)$, the margin function is defined as

$$mg(\boldsymbol{X},Y) = av_k I(h_k(\boldsymbol{X})=Y) - max_{j\neq Y}\, av_k I(h_k(\boldsymbol{X})= j) \ldots\ldots (1)$$

The margin function indicates the confidence in the classification by measuring the differences between the average number of votes at X, Y for the correct class and that for any other class. The generalization error is also defined as

$$PE^* = P_{\boldsymbol{X},Y}(mg(\boldsymbol{X},Y) < 0) \ldots\ldots\ldots \quad (2)$$

where $P_{\boldsymbol{X},Y}$ is the probability is over the **X**,Y region.

where $f_k(X) = f(X,\Theta_k)$, (Breimen, L, 2001) suggests that as the number of trees start increasing with strong correlation with the law of large numbers, certainly all sequences Θ_1…., *PE* will converge to*

$$P_{\boldsymbol{X},Y}(P_\Theta(h(X,\Theta)=Y) - max_{j\neq Y} P_\Theta(h(X,\Theta)= j) < 0)\ldots\ldots \quad (3)$$

The prediction system of random forest estimates the mean or mode of individual decision tree, with decision rules to estimate the result. Decision tree learning merits usages by many due to ease of interpretation of the process. Pavlov (2019) and Mitchell,

(2011) observed that random forest accuracy was as good and sometimes better than Adaboost; it is robust to outliers and noise; it is faster than bagging or boosting; it provides meaningful internal estimate of errors, correlation and variable importance and; it is simple and easy for parallel computing. Random forest ensemble classification has been applied in many fields with good results, ranging from health-care such as in the diagnose of alcohol use disorder (Ogretmen, 2019) with higher accurate result when some features with no apparent contribution to the accuracy are carefully eliminated than when all features were used. Fecal source identification was conducted by Roguet et al. (2018) with rapid and accurate solution using random forest. Long et al. (2019) applied random forest to analysis and prediction of travel mode choices by applying 2013 travel diary data from Nanjing, China. The method achieved high accuracy and also less computation cost. Random forest (RF), boosting and support vector machine (SVM) predictive accuracies have been compared by Ogutu et al (2011) for genomic selection. Random forest accuracy came second behind boosting. It surpassed the SVM and others with empirical evidence. Out-of-Bag (OOB) samples are samples not included in the training samples which are used for error check and accuracy assessment in random forest classifier. It is equally used for selection of appropriate values for tunning parameters with reduced bias, resulting in good estimation when stratified subsampling is applied (Janitza & Hornung, 2018) The algorithm is computationally cheaper than, and serves as alternative to cross-validation error metric (Park & Ho, 2018)

III. Methodology

Landsat satellite image data covering the area of interest (Aoi) with scene IDs as `LC81880552017006LGN01` and `LC81890552017045LGN00` were downloaded from the host, USGS website. The images were corrected for radiometric and atmospheric corrections through the use of ENVI image processing software. The two images were mosaicked and the Aoi was culled. The image bands were stacked. The stacked image contains seven bands: one to seven, excluding the band 8. Each image band is characterized with 30m-by-30m spatial resolution. The image was loaded as raster data by QGIS software and the training areas were selected. These areas were already known through visitation to site at earlier dates. They were saved as ESRI shapefile for later usage. The entire Aoi image was also saved with .img extension.

A. Study Area

The study area is a part of lower Benue trough located in Kogi state, Nigeria (Figure 1). The area can be defined by the following geographic coordinates framework.

Lower Left: 7.225164ºN, 6.865464ºE

Upper Right: 7.837803ºN, 7.841064ºE

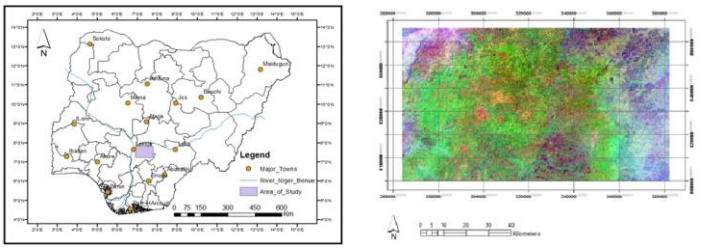

Figure 1. Left: Map of Nigeria Depicting Area of Study, Right: Processed Image of the Study Area.

B. Machine Learning procedure

The training data and the entire Aoi data were loaded using the ogr and the gdal of the osgeo package respectively. The training data were rasterized and classed into coal and non-coal areas. The array was observed to have conformity. The Aoi image was masked for cloud coverage and one of the bands was later excluded out of the seven. the entire bands were found to contain 8682030 pixels. While the Class 1 which represents the non-coal pixels in the training data has 15718 pixels, the coal area training data contains 510 pixels and classed as 2. The model was conducted with 500 decision trees with OOB score set as 'True', and the training data was fitted. Accuracy of prediction and confusion matrix were produced to verify the reliability of the model. The model was adopted to predict the entire Aoi image pixels. Visualization was conducted to see the results. Results were exported to file for further treatment with ArcMap. In ArcMap, overlay of the result was done on the bands 751 false colour RGB image of the Aoi in order to effectively compare the predicted areas with reserved validation areas.

IV. Results

Figure 2 depicts plotted short-wave infrared 1, SWIR1 band of the image and the region of interest training data. Areas encircled in black represent the non-coal area training sites while the areas encircled in blue are the areas earlier exploited for coal as identified on ground and google earth image (figure 3b). Areas encircled in red in figure 3b are other identified areas on ground where coal had been exploited but serve as reserved areas for validation.

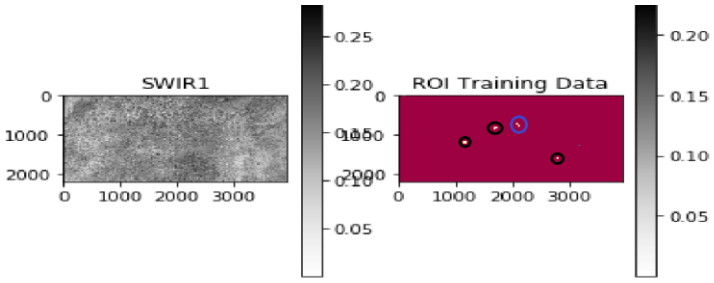

Figure 2. Left: SWIR1 Band and, Right: Training Data.

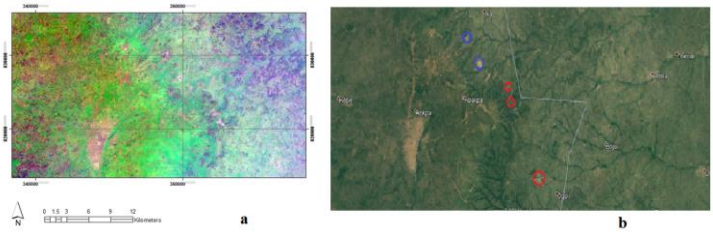

Figure 3. a: Landsat RGB 751 False Colour Composite of the entire Aoi. b: Google earth image of the AoI showing Coal Pits. Blue marker represents site used for training while red markers are validation pits.

Figure 4a shows the classified map. Bright areas represent the exploited areas for coal (coal pits) while other areas in dark hue are non-coal areas. Accuracy of prediction was 99.957%.

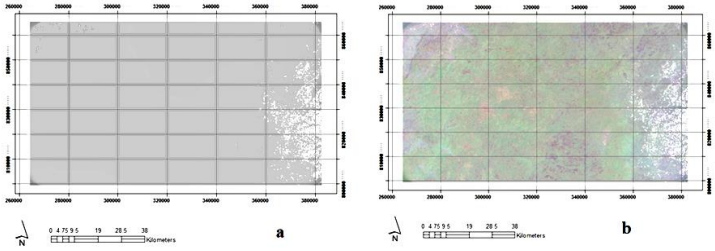

Figure 4. a: coal pit map. The Pits are Shown in Bright Pixels. b: Aoi Image Super-imposed on the Pit Map.

Figure 4b is the super-imposed 4a on the AoI image at 65 percent transparency of the latter. Figure 4c is the zoomed-in image of the portion of figure 4b which contained both the train data and the validation pit for better verification. The spatial coordinates of validation pits were further checked and found to correspond.

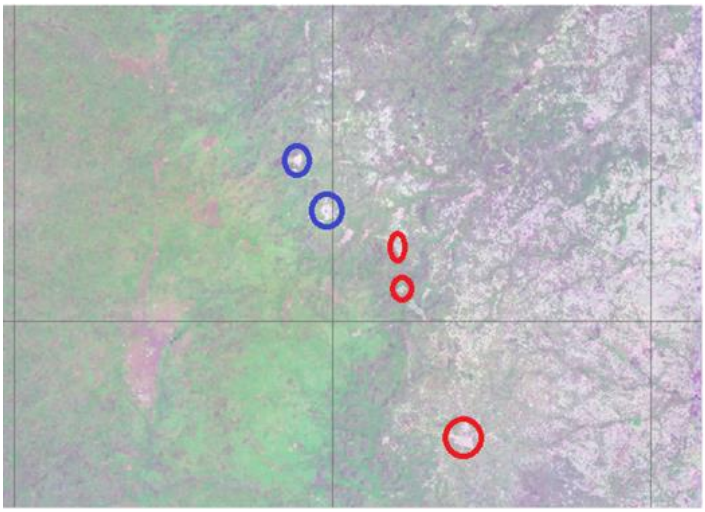

Figure 5. Zoomed-in Image of the Eastern Flank of Figure 4b for Verifying How well the Classified Pit and the Validation Pits Match.

V. Conclusion

The objective of this work was to map coal pits in part of Benue trough using random forest ensemble machine technique. This objective was achieved with very good accuracy of result. Checks made against reserved areas as used showed consistency. The method was able to detect other available pits on the eastern flank of the AoI.

References

[1] Abellera, R., Bulusu, L., Abellera, R., & Bulusu, L. (2018). "Machine Learning with OBIEE. In *Oracle Business Intelligence with Machine Learning",* https://doi.org/10.1007/978-1-4842-3255-2_4

[2] Albon, C. (2016). "*Python Machine Learning Cookbook",* (R. Roumeliotis & B. Jeff (eds.)). O'Relly Media inc. http://proquest.safaribooksonline.com.ezproxy.lib.vt.edu/9781786464477

[3] Breimen, L, (2001) "Random Forests", Statistics Department, University of California, Berkeley, CA. (Thesis).

[4] Famuboni, A. D. (1996). "Maximizing Exploration of Nigeria's Coal Reserves", In: Nigerian Coals: A Resource for Energy and Investment. *Raw Materials Research and DevelopmentCouncil (RMRDC)*, 39–62.

[5] Janitza, S., & Hornung, R. (2018). On the overestimation of random forest ' s out-of-bag error. *PLoS ONE*, *13*(7), 1–31

[6] Long, C., Xuewu, C., Jonas, D. vos, Xinjun, L., & Frank, W. (2019). "Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society"*, *14*, 1–10. https://doi.org/10.1016/j.tbs.2018.09.002

[7] Minjng. (2006, February). "Nigeria: An Exciting New Mining Destination". *Mining Journal Special Publication*, *Special Publication*, 1–20.

[8] Mitchel, M.W. (2001). Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters. Open Journal of Statistics, 2011(October), 205-211. https://doi.org/10.4236/ojs.2011.13024

[9] Ogretmen, B. (2019). HHS Public Access. "*Physiology & Behavior"*, *176*(3), 139–148. https://doi.org/10.1016/j.neulet.2018.04.007.Random

[10] Ogutu, J. O., Piepho, H. P., & Schulz-Streeck, T. (2011). "A comparison of random forests, boosting and support vector machines for genomic selection". *BMC Proceedings*, *5*(SUPPL. 3), 3–7. https://doi.org/10.1186/1753-6561-5-S3-S

[11] Park, Y., & Ho, J. C. (2018). PaloBoost: An Overfitting-robust TreeBoost with Out-of-Bag Sample Regularization Techniques. *Association for Computing Machinery*, *1*(1), 1–21

[12] Pavlov, Y. L. (2019). "Random forests. *Random Forests*," 1–122. https://doi.org/10.1201/9780367816377-11

[13] Roguet, A., Eren, A. M., Newton, R. J., & McLellan, S. L. (2018). "Fecal source identification using random forest". *Microbiome*, *6*(1), 1–15. https://doi.org/10.1186/s40168-018-0568-3

Authors

First Author – Okeke Francis Ifeanyi, PhD (Geodesy and Geoinformatics), University of Nigeria, Nsukka, francis.okeke@unn.edu.ng
Second Author – Ibrahim Adesina Adekunle, MTech (Remote Sensing), ibrahim.adekunle.pg.79358@unn.edu.ng
Third Author – Echeonwu Emmanuel Chinyere, MSc (Geospatial & Mapping Sciences), MSc (Machine Intelligence), African Institute for Mathematical Sciences, Rwanda, eecheonwu@aimsammi.org

Correspondence Author – Ibrahim Adesina Adekunle, ibrahim.adekunle.pg.79358@unn.edu.ng, adekunleibrahim6@gmail.com, +2348035086164

http://dx.doi.org/10.29322/IJSRP.10.12.2020.p10851 www.ijsrp.org