MAIZE YIELD ESTIMATION IN KENYA USING MODIS

B. K. Kenduiywo[a,b]*, A. Ghosh[b,c], R. Hijmans[b], and L. Ndungu[d]

[a]Department of Geomatic Engineering and Geospatial Information Systems,
Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya - bkenduiywo@jkuat.ac.ke
[b]Environmental Science and Policy, University of California, Davis, USA - (bkenduiywo,anighosh,rhijmans)@ucdavis.edu
[c]Alliance of Bioversity International and CIAT, Africa Hub, Nairobi, Kenya - a.ghosh@cgiar.org
[d]Regional Centre for Mapping of Resource for Development, Nairobi, Kenya - lndungu@rcmrd.org

Commission III, WG III/10

KEY WORDS: Maize, Kenya, Yield Estimation, MODIS, NDVI, GNDVI, GPP, FPAR

ABSTRACT:

Monitoring staple crop production can support agricultural research, business such as crop insurance, and government policy. Obtaining accurate estimates through field work is very expensive, and estimating it through remote sensing is promising. We estimated county-level maize yield for the 37 maize producing countries in Kenya from 2010 to 2017 using Moderate Resolution Imaging Spectroradiometer (MODIS) data. Support Vector Regression (SVR) and Random Forest (RF) were used to fit models with observed county level maize yield as a function of vegetation indices. The following five MODIS vegetation indices were used: green normalized difference vegetation index, normalized difference vegetation index, normalized difference moisture index, gross primary production, and fraction of photosynthetically active radiation. The models were evaluated with 5-fold leave one year out cross-validation. For SVR, R^2 was 0.70, the Root Mean Square Error (RMSE) was 0.50 MT/ha and Mean Absolute Percentage Error (MAPE) was 27.6%. On the other hand for RF these were 0.69, 0.51 MT/ha and 29.3% respectively. These results are promising and should be tested in specific applications to understand if they are good enough for use.

1. INTRODUCTION

In Kenya, crop production is a vital contributor to food security and employment. The sector directly accounts for about 26% and indirectly for another 25% of gross domestic product (Machado and Paglietti, 2015; Kenya National Bureau of Statistics, 2017). Maize is the main staple food in Kenya. Kenya has about 2.1 million ha of maize, more than 40% of the total cropland area. Maize yields are variable, as they are affected by droughts and pests. For example, Fall Army Worm infestations led to a drop in maize production by 6.3% in 2017 (Kenya National Bureau of Statistics, 2017) leading to a severe maize shortage. A quantitative and spatially-explicit understanding of variation in maize yield can support better investments, more efficient markets, and improved policy making. If yield estimates are timely, they can be used to avert food shortage through appropriate interventions such as imports.

Here we investigate the use of remote sensing vegetation metrics from 8-day Moderate Resolution Imaging Spectroradiometer (MODIS) products to estimate maize yields in Kenyan counties. We anticipate that remote sensing can provide cheap, early, and perhaps more accurate maize production estimates than the estimates based on ground based government surveys (Chivasa et al., 2017).

Remote sensing has been used to estimate crop yield with several regression like techniques. For example, ordinary least squares (Rojas, 2007; Kim et al., 2014), piecewise linear regression (Prasad et al., 2006), Back-propagation Neural Network (Panda et al., 2010), regression tree-based models (Johnson, 2014), empirical Leaf Area Index (LAI) regression model (Baez-Gonzalez et al., 2005), multiple linear regression and machine learning regression using Random Forest (RF) (Kim and Lee, 2016; Kayad et al., 2019; Sakamoto, 2020), SVR techniques (Kayad et al., 2019). Others have used convoultional neural networks (Kuwata and Shibasaki, 2015; Mu et al., 2019) or a combination of remote sensing and simulation models.

In the present study, we use SVR and RF regression models to predict maize yields based on MODIS vegetation indices in maize producing counties in Kenya. The counties were grouped into homogeneous regions with similar maize phenology. Only maize pixels extracted from an existing cropland map were used as described in Section 2.2. This cropland map was developed in 2015 using Landsat data. The process involved visual image interpretation by an analyst and guided on screen digitization. Therefore, vegetation indices derived from maize only pixels, were aggregated to county boundaries and used to model maize yields based on reference county level yields between 2010 and 2017. The reference yield data was obtained from the Kenya Ministry of Agriculture, Livestock, Fisheries and Irrigation (MOALFI).

The rest of the paper is organized as follows. Section 2 describes data used and explains the approach we adopted and illustrates how RS metrics from MODIS were used for maize yield prediction. In Section 3, the results are presented. This section is followed by the Discussion and Conclusions.

2. MATERIALS AND METHODS

2.1 Study area and data

Our study area encompasses the 37 Kenyan counties that grow maize. The counties are grouped into 8 regions with respect to similarity in the maize cropping calendar (Table 1, Figure 1). Trans Nzoia and Uasin Gishu counties are the major producers of maize in Kenya.

*Corresponding author.

We only considered the long-rain season and defined our start and end of growing season with the guide of regional maize calendar in (GEOGLAM, 2020) and Normalized Difference Vegetation Index (NDVI) onset and offset. Generally, we used the months of March–November in Coast and North Rift, April–November for South Rift and Central, March–September for upper and lower Eastern, March–October for upper Nyanza and the whole year for Western because of dominant double season maize growing.

Table 1. Grouped maize growing counties in Kenya.

Region	County names
North Rift	1. Baringo, 2. Nandi, 3. Uasin Gishu, 4. Trans Nzoia, and 5. Elgeyo Marakwet
South Rift	6. Bomet, 7. Kericho, 8. Nakuru, and 9. Narok
Central	10. Nyandarua, 11. Nyeri, 12. Kiambu, 13. Murang'a, and 14. Kirinyaga
Upper Eastern	15. Marsabit, 16. Tharaka Nithi, 17. Isiolo, 18. Meru, and 19. Embu
Coast	20. Taita Taveta, 21. Kwale, 22. Kilifi, 23. Tana River, and 24. Lamu
Upper Nyanza	25. Kisii and 26. Nyamira
Western	27. Migori, 28. Homa Bay, 29. Kisumu, 30. Siaya, 31. Kakamega, 32. Bungoma, 33. Busia, and 34. Vihiga
Lower Eastern	35. Makueni, 36. Kitui, and 37. Machakos

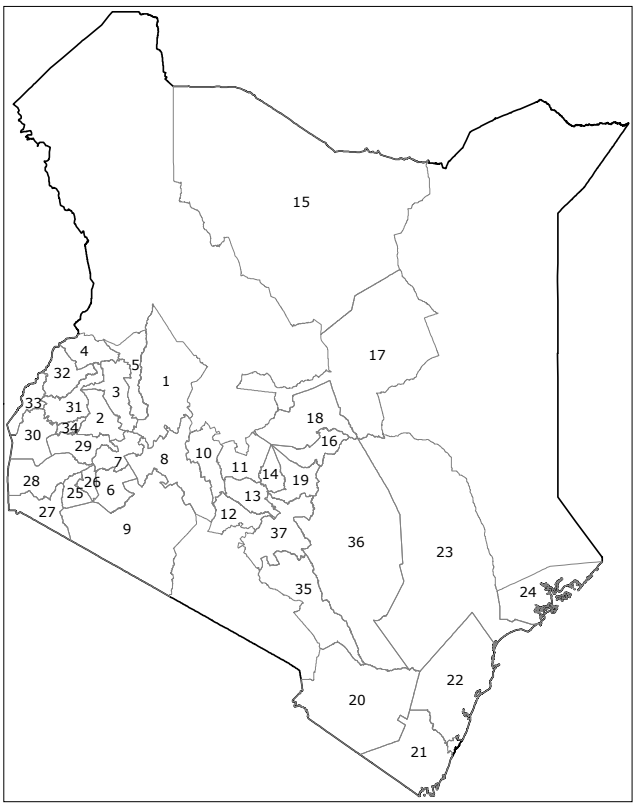

Figure 1. Maize growing counties in Kenya used for yield prediction. The number labels correspond to County names in Table 1.

County level maize yield data for 2010 to 2017 was obtained from the MOALFI which has made it available via (MOALF, 2020) (Figure 2). The data was collected by the Kenyan government field extension officials under the state department of agriculture. The data is being continuously made available online through Global Open Data for Agriculture and Nutrition initiative. There are clear regional differences in maize yield, with the highest yield in the North Rift region, followed by South Rift, Nyanza, Western, Central, Coast, upper Eastern and lower Eastern.

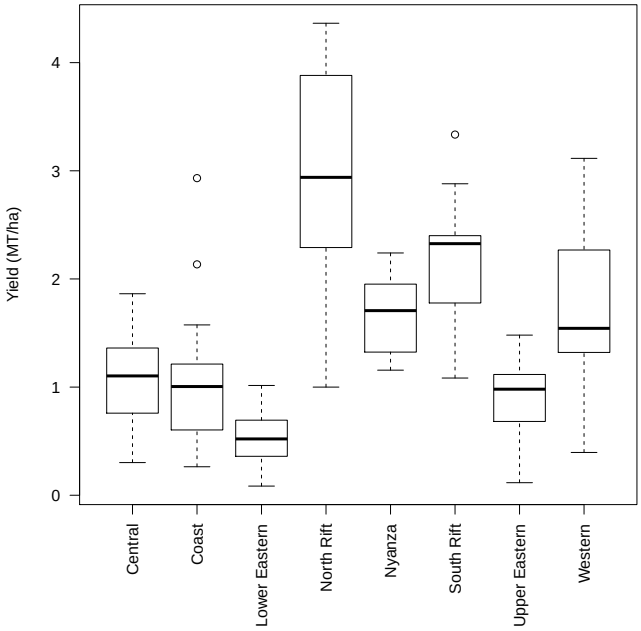

Figure 2. Regional distribution of maize yield data acquired between 2010–2017.

2.2 Methods

Figure 3 gives an overview of the techniques used to implement this study. Basic tasks of our methodological framework includes MODIS data acquisition, RS metrics computation, masking out non-maize areas using maize maps, exclusion of atmospheric and sensor affected pixels using MODIS quality masks, aggregation of the metrics spatially and temporal per county, maize yield prediction using SVR and RF, and validation of model predictions. Details of these steps are described in subsequent subsections.

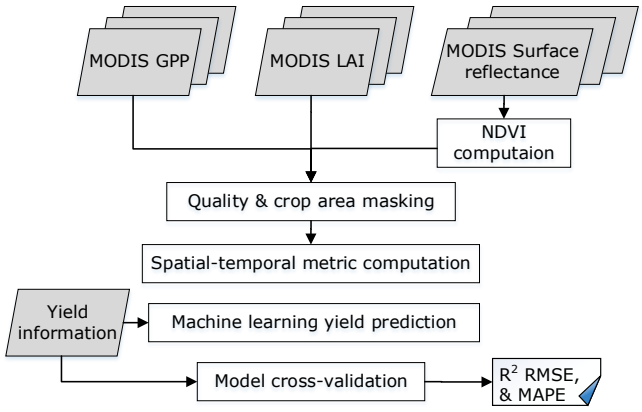

Figure 3. Methodological framework adopted for maize yield prediction.

2.2.1 MODIS data processing To predict yield, we used the following MODIS data products: (1) NDVI (2) Green Normalized difference Vegetation Index (GNDVI), (3) Leaf Area Index (LAI),

(4) Gross Primary Productivity (GPP), (5) Normalized Difference Moisture Index (NDMI), and (6) Fraction of Photosynthetically Active Radiation (FPAR). NDVI, GNDVI, and NDMI were computed from MODIS 8-day 500 m surface reflectance data found in the MOD09 series products from the Terra satellite (NASA, 2020). The NDVI is commonly used as a proxy for green biomass. It is computed as the ratio of the reflectance in the near infra-red (NIR) and red portion of electromagnetic spectrum, that is,

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}. \quad (1)$$

The GNDVI substitutes the red band in NDVI equation with green as

$$\text{GNDVI} = \frac{\text{NIR} - \text{GREEN}}{\text{NIR} + \text{GREEN}}. \quad (2)$$

The GNDVI was developed to estimate chlorophyll concentration in vegetation (Gitelson et al., 1996) and may be useful as a proxy for photosynthetic rate and plant stress. The NDMI is given as

$$\text{NDMI} = \frac{\text{NIR} - \text{SWIR1}}{\text{NIR} + \text{SWIR1}} \quad (3)$$

where *SWIR1* is the short wave infra-red 1 band in MODIS surface reflectance. We used it in order to quantify water content in maize since it is sensitive to the moisture levels in vegetation. Basically, soil moisture variability is one of the main factors affecting crops productivity. Lastly, GPP, LAI and FPAR metrics are 8-day 500 m products from MODIS. GPP is a product from MODIS that was acquired from MOD17 data series products generated from Terra satellite. It is based on the radiation-use efficiency concept and can potentially be used to quantify generation of new biomass in vegetation. The LAI and FPAR are found in MOD15 product series of MODIS. LAI is a one-sided green leaf area per unit ground surface area dimensionless quantity that characterizes plant canopies. In contrast, FPAR is an important parameter in estimating biomass production because the development of vegetation is related to the rate at which radiant energy is absorbed by vegetation. Compared to NDVI, GPP, LAI and FPAR model-based biophysical variables normally show good correlation with crop yield and primary production (Coleman et al., 2017).

After computing the RS metrics, we masked out atmospheric effects, water and data affected by varying sensor conditions using MODIS quality masks that come with the products. For instance, we masked out pixels with clouds, shadows, water areas, aerosol, cirrus, fire and snow from MOD09 surface reflectance product. In LAI and FPAR products, pixels with water, snow, aerosol, cirrus, and shadows were masked out. Similarly, pixels with clouds, dead detector, and with poor confidence quality score were excluded. Finally, a second crop mask was applied on quality masked image scenes in order to retain maize growing areas only within each county.

The masked images were used to compute spatial-temporal metrics for each county using the process summarized in Figure 4. This was done by first computing mean aggregates of all pixels within each county boundary for each image scene to obtain spatial metrics. A mean aggregate of all the spatial metrics within a defined maize season was finally computed to obtain spatial-temporal metrics. This procedure is available on Earth Engine: https://code.earthengine.google.com/60abb28e6af6e56296452591192e1e5e.

2.2.2 Feature selection Feature selection is a process of selecting relevant variables that aid model prediction. It is an important step that helps minimize model over-fitting while aiding its prediction accuracy. We used RF's mean decrease in accuracy measure from variable importance to select relevant metrics from the initial 6 that were computed. In principle, mean decrease in accuracy is computed by determining the impact a predicting variable has when it is removed from the model. Figure 5 shows the outcome of RF feature importance. Basically, GPP was the most important metric in maize yield prediction followed by NDVI, FPAR, LAI, NDMI and GNDVI. Following this guide we selected all variables except LAI with consideration of information diversity.

2.3 Maize yield prediction

We tested two machine learning methods, RF and SVR, for maize yield prediction using the RS metrics selected earlier. These models were adopted because previous studies have shown that they lead to good results compared to other methods (Kim and Lee, 2016; Kayad et al., 2019; Sakamoto, 2020).

2.3.1 Random Forest (RF) machine learning ensemble technique is based on CART (Classification and Regression Trees) (Breiman, 2001). Random forest fits many trees with a bootstrapped sample, and also takes a random sample of the variables that can be used at each split in the tree (James et al., 2013). We set the number of trees to 500 and the number of variables used to split nodes as n/3, where n = number of input variables.

2.3.2 Support Vector Regression (SVR) Support vector machines has gained popularity in image classification and regression (Vapnik, 2000). SVR is a generalization of the classification problem where the model returns a continuous-valued output as opposed to an output from a finite set. Predictions are done in SVR by using an optimal hyperplane to minimizes prediction error. We used radial basis kernel to construct the model's hyperplane. The kernel has two parameters namely ϵ and penalty parameter C. We determined these parameters via a grid search based on the least mean square error.

2.3.3 Model evaluation We used cross-validation to compute Root Mean Square Error, Mean Absolute Percentage Error (MAPE) and coefficient of determination R^2 measures from data with a pair of yield and corresponding RS metrics. For example, given observed yields y and their corresponding predicted yields $\hat{y}$, the RMSE is computed as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{n}}, \quad (4)$$

while MAPE is

$$\text{MAPE} = \frac{\frac{\sum_{i=1}^{n}|y_i - \hat{y_i}|}{y_i}}{n} \times 100\%, \quad (5)$$

and

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \hat{y_i}/n)^2} \quad (6)$$

where n is the number of observations. The smaller the RMSE value, the closer are predicted maize yields to actual ones. MAPE is an average of the absolute percentage errors from model predictions, i.e., an average of the ratio of absolute yield errors with actual yields (Equation (5)). This measure expresses prediction error as a percentage allowing for comparisons between studies. Lastly, the R^2 explains the proportion of variance in the dependent variable that is explained by the independent variable. We used 5-fold leave one year out cross validation to compute these model evaluation measures.

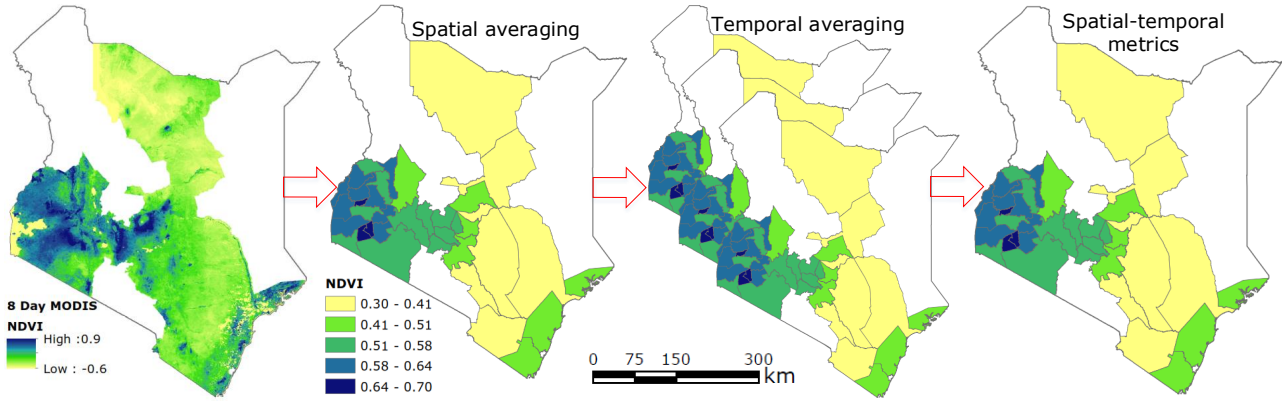

Figure 4. Remote sensing processing steps in GEE.

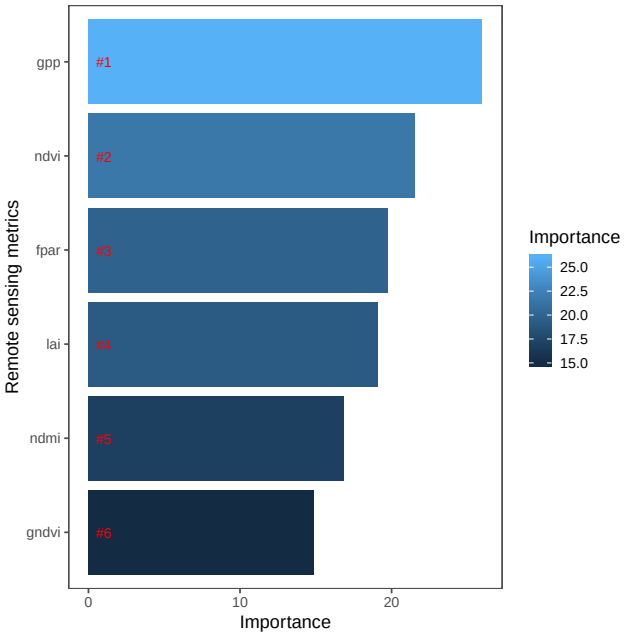

Figure 5. Selection of remote sensing yield prediction metrics using RF.

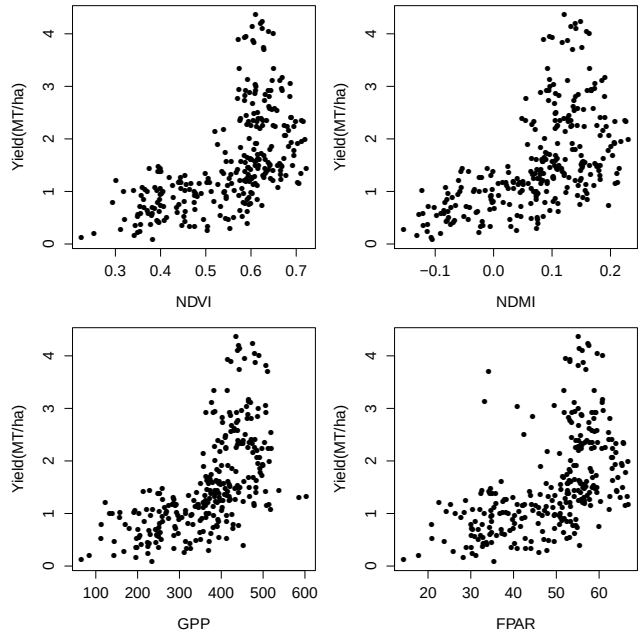

Figure 6. Relationship of RS metrics with maize yields over 2010–2017 period.

3. RESULTS

Figure 6 shows some of the metrics to used predict maize yields. All the metrics show an asymptotic relationship with maize yields. Maize yields increase linearly with NDVI from around 0.1 to 0.5 which corresponds to maize yields between 0–2 ton/ha. From 0.5 NDVI rises sharply to 0.7 which corresponds to yields between 2–5 ton/ha. In NDMI, a linear relationship is depicted between -0.1 to 0.1 are consistent with maize yield between 0–2 ton/ha like NDVI. When NDMI is in the range of 0.1–0.25 the maize yields sharply increase between 2–5 ton/ha. GPP exhibits a relationship with maize yields with values of ranges 100–500. Lastly, FPAR shows a relationship with maize yields when it ranges between 10–60 though with some outliers.

Cross validation results are shown in Figures 7 and 8. SVR had a RMSE 0.50 ton/ha, MAPE of 27.6% and R^2 of 0.7 which was slightly better than the results obtained with RF.

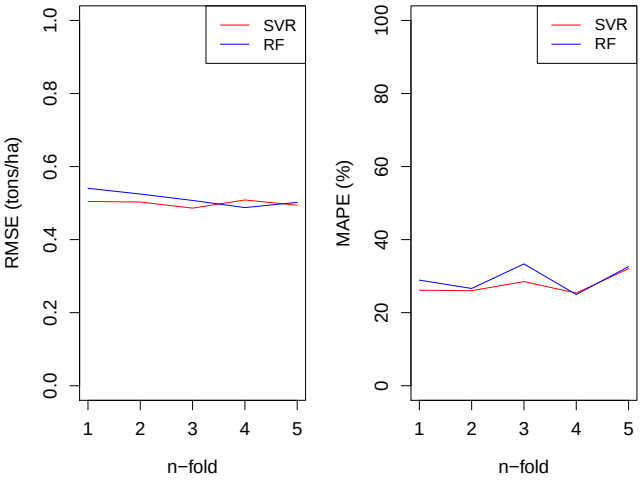

Figure 7. RMSE and MAPE from 5-fold cross-validation of RF and SVR maize prediction models.

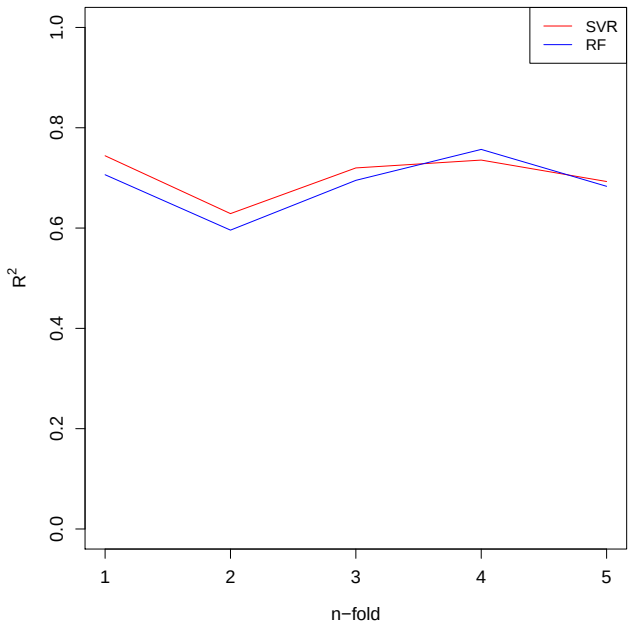

Figure 8. Coefficient of determination R^2 of RF and SVM for a 5-fold cross-validation.

4. DISCUSSION

We have adopted machine learning regressing techniques to predict maize yields in Kenya using MOALFI data collected annually. The objective is to provide a remotely sensed platform for rapid yield estimation during maize growing season. Maize is a staple food for most Kenyan families and is also a source of income. Due to lack of proper maize estimates farmers have suffered from poor maize prices and other times shortage during low seasons that results to food scarcity. Therefore we adopted RS metrics from MODIS satellite for yield prediction. All the metrics are correlated to county yields recorded between 2010–2017 (Figure 6). The GPP metric during maize growing season had the highest feature importance (Figure 5). This is expected because GPP acquired during growing season period has been established to be one of the best indicators of the amount of new biomass (Prince, 1991; Gitelson et al., 2006) in crops and hence the reason it correlates well with maize yields. In contrast to findings by (Shanahan et al., 2001), which demonstrated that GNDVI acquired during mid-grain season was the most highly correlated with grain yield, GNDVI had the lowest importance in our study. This is because our study used GNDVI mean aggregate from the entire season as opposed to mid-gran period only.

Selected (NDVI, GNDVI, NDMI, GPP, and FPAR) metrics were used to predict maize yields using SVR and RF machine learning methods. The performance of SVR and RF was very similar. Both methods explained a large amount of yield variability. We established that the RMSE of 0.50 ton/ha (SVR) and 0.51 ton/ha (RF) is an improvement over other studies like (Guindin-Garcia, 2010). The average predictor error attained by the two approaches, i.e. 27.6% in SVR and 29.3% in RF, may be sufficiently accurate for use; but it is also clear that there is much room for improvement.

Our study has demonstrated that it is possible to predict maize yields in Kenya using MOALFI historical data. Despite these encouraging findings there is still more room to improve yield predictions. For instance, we used a maize crop mask that was generated in 2015 via expert knowledge digitization. We expect that there may have been changes in maize growing area in different counties between 2010-2017 period that we used for model prediction. Though we assumed, such changes to be negligible, use of maize crop mask generated annually to compute RS metrics might improve prediction accuracy. This is a subject of our future study. It is also important to note that administrative boundaries have changed over time through different Kenya government regimes. These changes might have introduced biases while streamlining collected maize yield data from old to new administrative boundaries. Nonetheless, despite aggregating RS metrics to the county boundaries the prediction accuracy attained is reasonable. However, although RS data is increasingly accessible at better spatial-temporal resolution and at no cost, ground reference data is still essential to design and validate RS metrics based predictions (Coleman et al., 2017).

5. CONCLUSION AND OUTLOOK

The study has demonstrated that maize yield estimation in Kenya can be achieved at reasonable prediction accuracy using machine learning SVR and RF. Maize yield prediction can help MOALFI, traders and other food security stakeholders. In future work, we will consider regions with similar agro-ecological and cultural farming attributes and use annual maize mask generated by deep learning in our model predictions. We hope to design the models to predict yields at pixel level in each county.

ACKNOWLEDGEMENTS

This research is part of the Quality Index Insurance Certification (QUIIC) project that is funded by United States Agency for International Development (USAID) through University of California Davis in partnership with Regional Center for Mapping of Resources for Development (RCMRD) in Kenya.

REFERENCES

Baez-Gonzalez, A. D., Kiniry, J. R., Maas, S. J., Tiscareno, M. L., Macias, C. J., Mendoza, J. L., Richardson, C. W., Salinas, G. J. and Manjarrez, J. R., 2005. Large-Area Maize Yield Forecasting Using Leaf Area Index Based Yield Model. *Agronomy Journal* 97(2), pp. 418–425.

Breiman, L., 2001. Random forests. *Machine Learning* 45(1), pp. 5–32.

Chivasa, W., Mutanga, O. and Biradar, C., 2017. Application of remote sensing in estimating maize grain yield in heterogeneous African agricultural landscapes: a review. *International Journal of Remote Sensing* 38(23), pp. 6816–6845.

Coleman, E., Dick, W., Gilliams, S., Piccard, I., Rispoli, F. and Stoppa, A., 2017. Remote sensing for index insurance: Findings and lessons learned for smallholder agriculture. Technical report, International Fund of Agricultural Development (IFAD).

GEOGLAM, 2020. Crop Monitor. https://cropmonitor.org/index.php/eodatatools/cmet/. Accessed: 22nd February 2020.

Gitelson, A. A., Kaufman, Y. J. and Merzlyak, M. N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment* 58(3), pp. 289–298.

Gitelson, A. A., Viña, A., Verma, S. B., Rundquist, D. C., Arkebauer, T. J., Keydan, G., Leavitt, B., Ciganda, V., Burba, G. G. and Suyker, A. E., 2006. Relationship between gross primary production and chlorophyll content in crops: Implications for the synoptic monitoring of vegetation productivity. *Journal of Geophysical Research: Atmospheres* 111(D8), pp. 1–13.

Guindin-Garcia, N., 2010. Estimating Maize Grain Yield From Crop Biophysical Parameters Using Remote Sensing. PhD thesis, University of Nebraska, Lincoln, Nebraska.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning*. Vol. 112, Springer.

Johnson, D. M., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment* 141, pp. 116–128.

Kayad, A., Sozzi, M., Gatto, S., Marinello, F. and Pirotti, F., 2019. Monitoring Within-Field Variability of Corn Yield using Sentinel-2 and Machine Learning Techniques. *Remote Sensing* 11(2873), pp. 1–20.

Kenya National Bureau of Statistics, 2017. *Economic Survey 2017*. Kenya National Bureau of Statistics.

Kim, N. and Lee, Y., 2016. Machine Learning Approaches to Corn Yield Estimation Using Satellite Images and Climate Data: A Case of Iowa State. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*.

Kim, N., Cho, J., Shibasaki, R. and Lee, Y., 2014. Estimation of corn and soybeans yields of the us midwest using satellite imagery and climate dataset. *Journal of Climate Research* 9, pp. 315–329.

Kuwata, K. and Shibasaki, R., 2015. Estimating crop yields with deep learning and remotely sensed data. In: *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 858–861.

Machado, M. D. and Paglietti, L., 2015. Kenya: irrigation market brief.

MOALF, 2020. Kilimo Open data. http://kilimodata.developlocal.org/dataset. Accessed: 22nd February 2020.

Mu, H., Zhou, L., Dang, X. and Yuan, B., 2019. Winter wheat yield estimation from multitemporal remote sensing images based on convolutional neural networks. In: *2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, pp. 1–4.

NASA, 2020. Land Processes Distributed Active Archive Center (LP DAAC). https://lpdaac.usgs.gov/. Accessed: 22nd February 2020.

Panda, S. S., Ames, D. P. and Panigrahi, S., 2010. Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques. *Remote Sensing* 2(3), pp. 673–696.

Prasad, A. K., Chai, L., Singh, R. P. and Kafatos, M., 2006. Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation* 8(1), pp. 26–33.

Prince, S. D., 1991. A model of regional primary production for use with coarse resolution satellite data. *International Journal of Remote Sensing* 12(6), pp. 1313–1330.

Rojas, O., 2007. Operational maize yield model development and validation based on remote sensing and agrometeorological data in Kenya. *International Journal of Remote Sensing* 28(17), pp. 3775–3793.

Sakamoto, T., 2020. Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing* 160, pp. 208–228.

Shanahan, J. F., Schepers, J. S., Francis, D. D., Varvel, G. E., Wilhelm, W. W., Tringe, J. M., Schlemmer, M. R. and Major, D. J., 2001. Use of remotesensing imagery to estimate corn grain yield. *Agronomy Journal* 93(3), pp. 583–589.

Vapnik, V. N., 2000. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.