Using crowdsourcing and machine learning for predicting the spatial distribution of banana-based cropping systems in Uganda

Dennis Ochola[1,2]*, Godfrey Taulya[2], Gerrie van de Ven[1], Ken Giller[1]

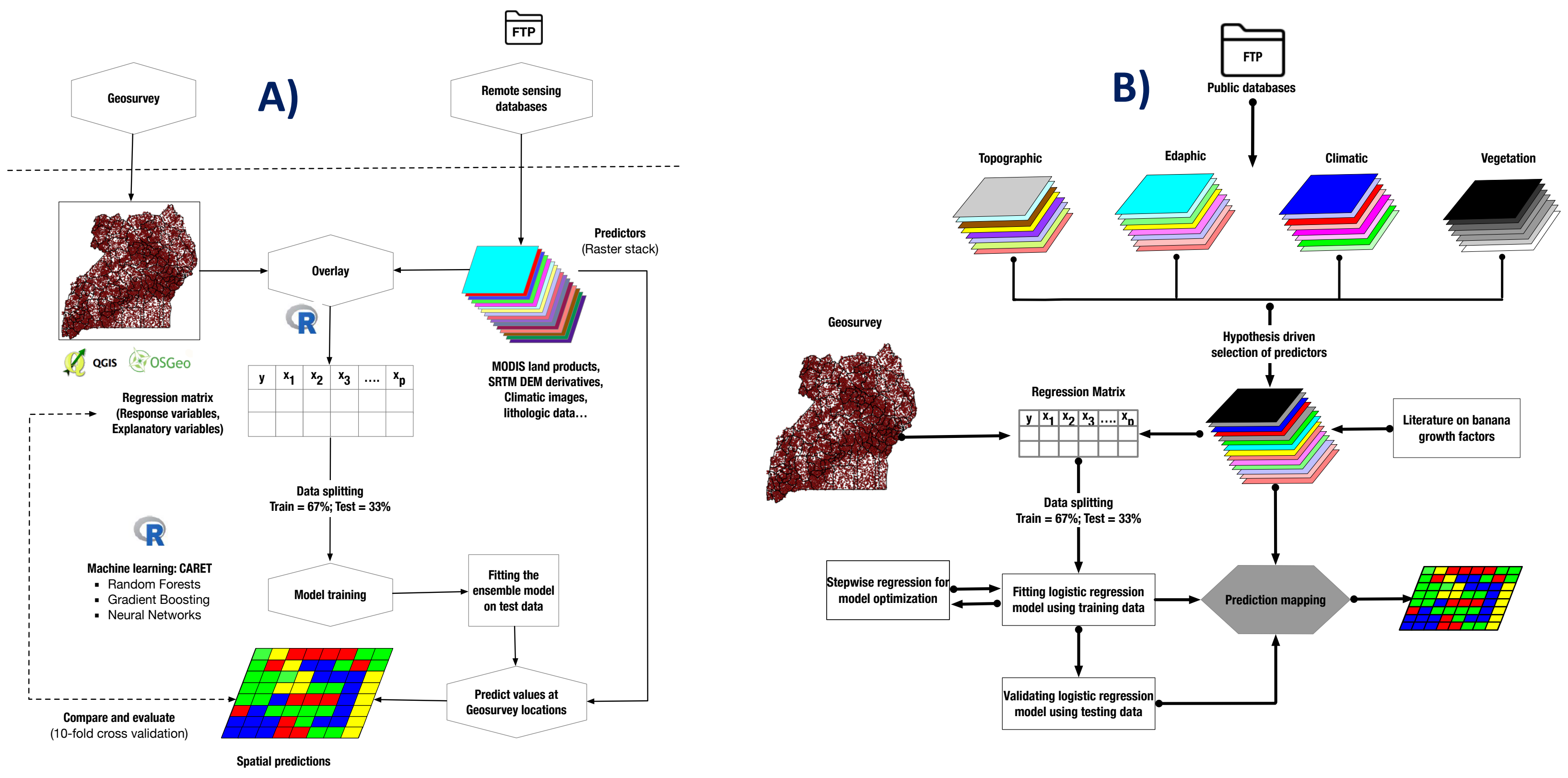
[1]Plant Production Systems, Wageningen University & Research (WUR), The Netherlands
[2]International Institute of Tropical Agriculture (IITA), Uganda

Introduction

Expert opinion has for decades been the basis of information on the distribution of banana-based cropping systems in Uganda, and elsewhere in East Africa. Lack of accurate and reliable spatial data undermines strategic planning and sustainable intensification at various scales. Few studies (e.g. Eledu et al., 2004) have attempted to identify and map the principal banana growing areas in East Africa. Hence, this study compares the prediction accuracy of machine learning and logistic regression, and applies the best approach to provide insights on the geographic shifts of banana production from 1958-2016.

Data processing and spatial prediction

18,959 presence and absence data were coupled with 71 covariates (21 climatic, 19 edaphic, 19 vegetation, 6 topographic and 6 socioeconomic) and split into 67% training and 33% testing datasets. Machine learning predictions using Random Forest (RF), Gradient Boosting Machines (GBM) and Neural Networks (NNET) (Fig 1A). Logistic regression mapping with 12 covariates (5 climatic, 6 edaphic and 1 vegetative) known to influence banana growth and physiology (i.e. hypothesis-based selection) (Fig 1B).



A)

FTP

Geosurvey

Remote sensing databases

Overlay

Predictors (Raster stack)

QGIS OSGeo

Regression matrix (Response variables, Explanatory variables)

y	x_1	x_2	x_3	….	x_p

MODIS land products, SRTM DEM derivatives, Climatic images, lithologic data…

Data splitting
Train = 67%; Test = 33%

Machine learning: CARET
- Random Forests
- Gradient Boosting
- Neural Networks

Model training

Fitting the ensemble model on test data

Predict values at Geosurvey locations

Compare and evaluate (10-fold cross validation)

Spatial predictions

B)

FTP

Public databases

Topographic

Edaphic

Climatic

Vegetation

Geosurvey

Hypothesis driven selection of predictors

Regression Matrix

y	x_1	x_2	x_3	….	x_p

Literature on banana growth factors

Data splitting
Train = 67%; Test = 33%

Stepwise regression for model optimization

Fitting logistic regression model using training data

Prediction mapping

Validating logistic regression model using testing data

Fig 1: *Workflows A) machine learning, B) logistic regression*

Machine learning vs. logistic regression

Algorithms RF and GBM performed better than NNET in terms of accuracy, receiver operating characteristic (ROC) and sensitivity. However, NNET performed better with regards to kappa and specificity. The ensemble model aggregating the prediction outcomes of RF, GBM and NNET performed better (AUC = 0.881) compared to the logistic regression model (AUC = 0.852) but not significantly different (p >= 0.05) (Fig 2). Logistic regression revealed that annual mean temperature, precipitation seasonality and cation exchange capacity (CEC) negatively influence spatial distribution of banana-based cropping systems.

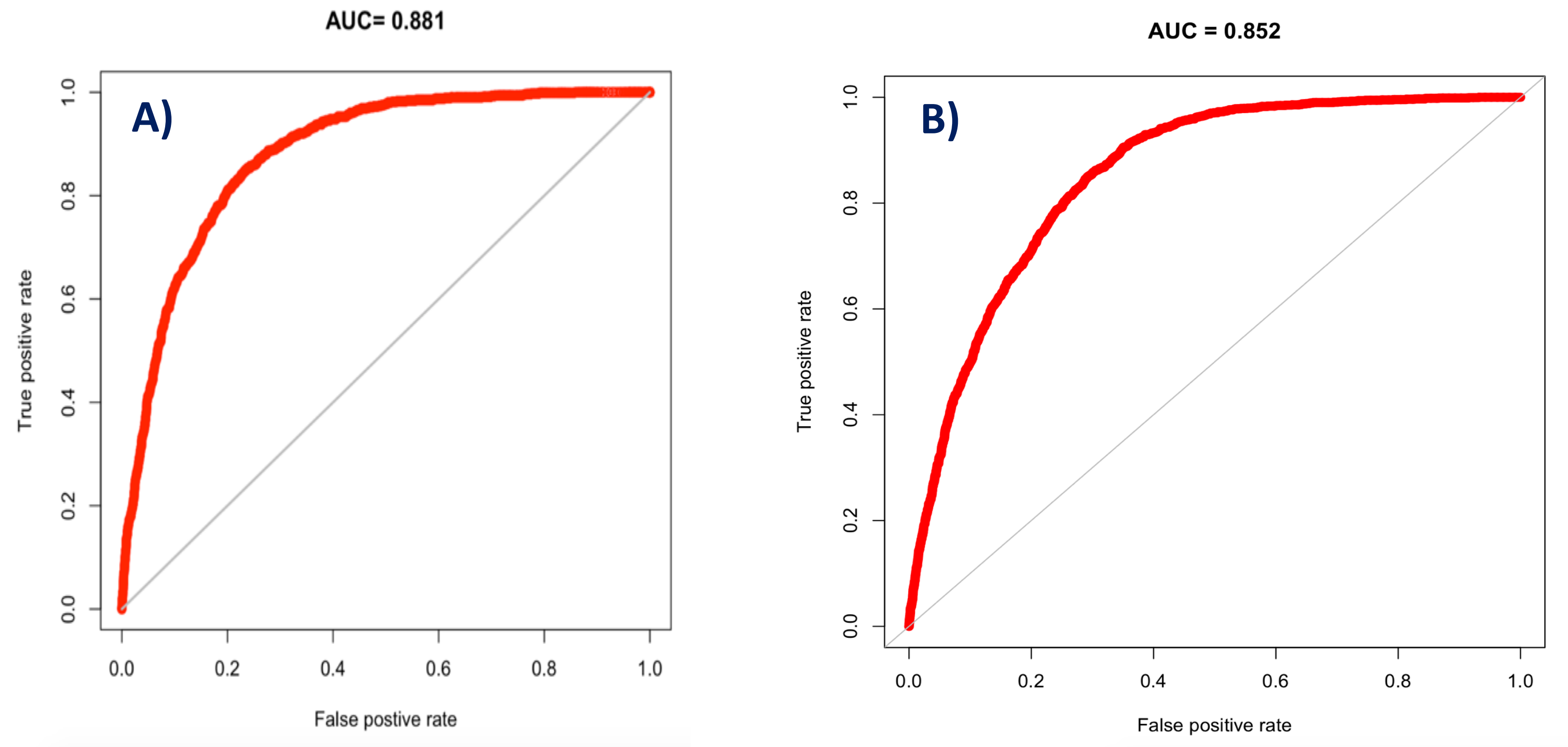


Fig 2: *Area under the ROC curve. A) ensemble model with 71 covariates, B) logistic regression model fitted with 12 selected variables*

Spatial distribution of bananas

Spatial prediction with the ensemble model (Fig 3A) reveal high probability of banana presence in the western (i.e. Ankole, Toro and foothills of Mt Rwenzori), central (i.e. Buganda in Kooki and Buddu) and in the eastern (i.e. foot hills of Mt Elgon) and least in the northern. Banana-based cropping systems occupied 9.6% of the land area of Uganda (Fig 3A). The kappa threshold of 0.249 slightly (-0.15%) underestimated distribution (Fig 3B).

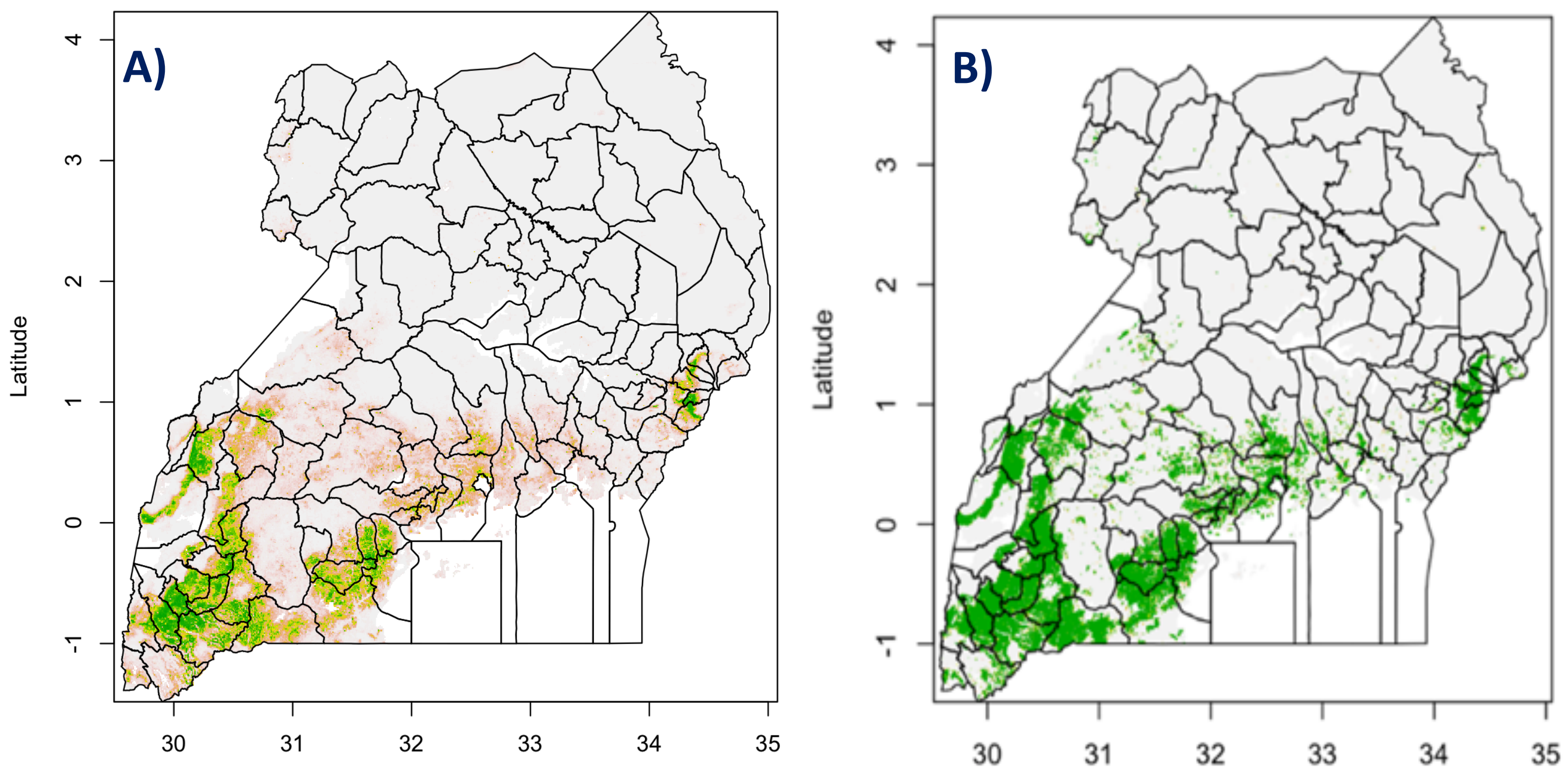


Fig 3: *Probability of the presence of banana-based cropping systems in Uganda A) machine learning, B) locations above the kappa threshold of 0.249*

Geographic shifts 1958-2016

- **1958:** Central (40.6%); Western (29.1%); Eastern (27.3%); Northern (3%) (Fig 4A).
- **2016:** Western (46.3%); Central (36.3%); Eastern (13.6%); Northern (3.4%) (Fig 4A).
- Geographic shifts defined mainly by areas where banana has shrunk (15%), expanded (41%) and stagnated (44%) (Fig 4B).
- Stagnation mainly in Central (-4.3%), expansion in Western (+17.3%) and shrinking in Eastern (-13.4%) (Fig 4B).

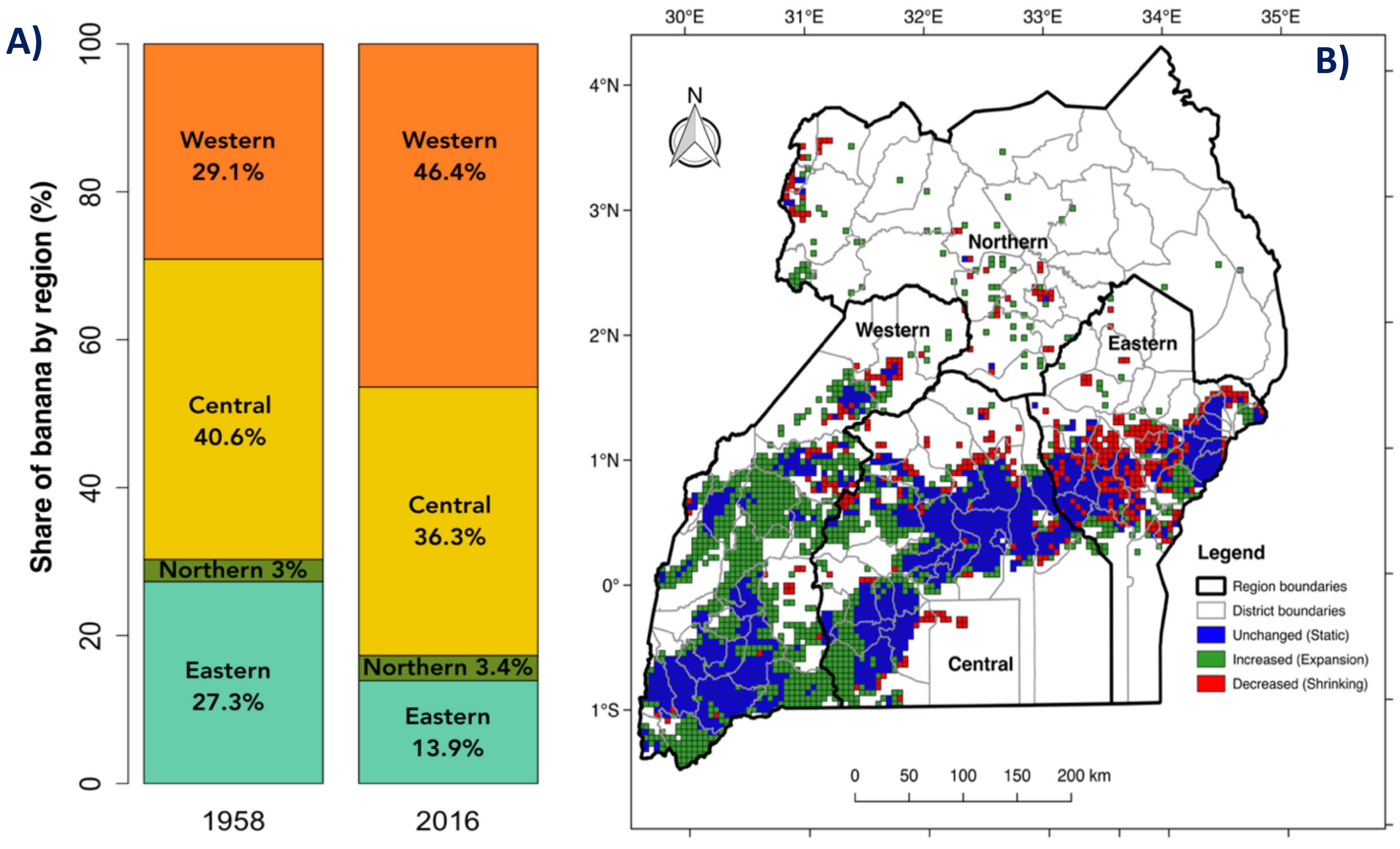


Fig 4: *A) Share of banana-based systems by regions, B) geographic shifts of banana production in Uganda from 1958 to 2016.*

Conclusion and way forward

- Machine learning can iteratively search and filter covariates to achieve high prediction accuracy, but inclusion of redundant covariates doesn't facilitate explicit description of outcomes.
- Hypothesis-based selection of covariates with known influence on banana growth and agronomic management is a better option for identifying drivers of geographical shifts.
- Mean annual temperature, precipitation seasonality and CEC negatively influence spatial distribution of banana-based cropping systems.

*Correspondence: dennis.ochola@wur.nl

Acknowledgment: Bill & Melinda Gates Foundation for funding the Banana Agronomy Project in Uganda