004

AKADEMIYA

The Expertise We Need. The Africa We Want.

AKADEMIYA2063 - *August 17 /August 24 -2020*

Covid-19 Bulletin

Application of Remote Sensing and Machine Learning for Crop Production Forecasting During Crises

Racine Ly, Director, Data Management, Digital Products and Technology;
Khadim Dia, Associate Data Scientist.

To mitigate the spread and impact of COVID-19, countries around the world have been taking policy action such as social distancing and mobility restrictions, as well as the closure of schools, businesses and national borders, among other measures.

Introduction and Context

The impact of these policies has put a heavy strain on different sectors of the economy and communities' livelihoods. One of the biggest fears is that COVID-19 and the public health measures taken also impact agricultural food production, leading to a potential food crisis in many parts of the world, especially in Africa.

While on the health side, official data shows that Africa is currently less affected by the COVID-19 pandemic compared to the rest of the world, there have been many impacts on African economies, which are largely dominated by the informal sector and as such, greatly dependent on the mobility of its populace. In the Agricultural sector, some undesired consequences on food crop production are inputs scarcity, the shortage of agricultural workers, access to export markets and food supply availability issues, including from imports. Good planning - from governments, policy and other decision makers - is crucial to anticipate and mitigate the potentially negative effects on the agricultural sector in order to prevent a food crisis. Being able to assess early how much food is expected to be produced provides not only a better overview of food security but also allows for more precise and disaggregated national food balance sheets.

African countries regularly issue production forecasts, generally at a national level, with limited disaggregation. The methodology used routinely relies on data samples collected at traditionally designated agricultural regions. The samples are then used for regression analysis to estimate production for other areas and/or at a national level.

FINDING A SOLUTION TO THE PROBLEM OF FOOD CROP PRODUCTION DATA

From effects on access to seeds and fertilizers, limited movement of goods, declining demand, to labor shortage, the disruptive impact of Covid-19 on food production systems is real. The challenge here is not only the likely extent and complexity of the disruptions but also the difficulty to identify and track them in real time. Unlike the propagation of the disease itself which can be tracked through testing and tracing, it is impossible, even in normal times, to have accurate information on cropping activities. The introduction of confinement and other measures to control the pandemic make the situation even more difficult. There is no way of knowing whether farmers have access to inputs, in time or in adequate quantities, whether they have been too sick to tend to their farmers or could work only partially. One would eventually find out at the end of the growing season from the impact of harvested quantities. One is then left to play catch up to deal with a crisis situation.

The complete lack of information about growing conditions can be overcome by using today's digital technologies. Remotely sensed data allow to track in real time changes in vegetation cover, weather data and other parameters related to cropping activities. Recent developments in machine learning and computer modeling make it possible to track and predict crop production using these data. The benefits go far beyond the ability to overcome the obstacles to data gathering during crises. The many weaknesses hampering the access to good quality agricultural statistics also can be overcome using the same digital technologies, from measuring arable land, planted areas, crop yields to the spatial distribution of harvested quantities. Our scientists are using these technologies to assess changes in food production systems during the pandemic and thereby provide valuable information to tackle the impact of the pandemic among local communities.

Ousmane Badiane, Executive Chairperson

Aside from involving costly and time-consuming data collection surveys, such techniques are known to lead to biased outcomes by not considering new agricultural lands or regions. Lastly, the same methods of forecasting food crop production currently fail to take into account more frequently occurring climate related shocks such as drought, floods, heat waves and pest invasions.

The multiple restrictions to control the pandemic add another layer of complications, as they make it difficult, at least when using traditional methods, to access the necessary data to evaluate the growing season early enough to be prepared for eventual negative outcomes. Moreover, the complex web of disruptions resulting from the pandemic renders the application of traditional forecasting techniques, which are not suitable for sudden shocks, impractical. Our team of scientists at AKADEMIYA2063 has developed a model to go around these various obstacles. The model uses remote sensing data and machine learning techniques to generate maps of predicted production, at fairly disaggregated geographic level, as an output. Thanks to *recent* developments in remote sensing, satellite images are able to observe features on earth in different wavelengths and with good temporal and spatial resolutions. They can revisit the same region of interest several times in a month and can cover a large region - even an entire country - which enhances our land monitoring capacities including agricultural ones. On the other hand, machine learning techniques recently gained attention due to the capacity to learning patterns within datasets without being explicitly programmed. We are applying the model to forecast the production of most food staple crops across Africa as the Covid-19 pandemic evolves in order to allow countries to anticipate and better target interventions to protect vulnerable groups and communities.

Solving data access problems through remote sensing and third-party maps

Production estimation based on remote sensing data can be done with two main approaches: (i) Using remotely sensed data as inputs to agro-meteorological or plant-physiological models, and (ii) Building a direct mathematical relationship between remotely sensed data and crop production (Huang & Han, 2014). The first approach is based on "mechanistic" descriptions of crop growth, development and production simulated through mathematical functions. Such methods have shown good results but are not able to exploit datasets to their full extent due to the constraints coming along the way crop growth phenomenon are described. The second approach usually relies on derived indicators from remotely sensed data and their correlation with crop growth and yield.

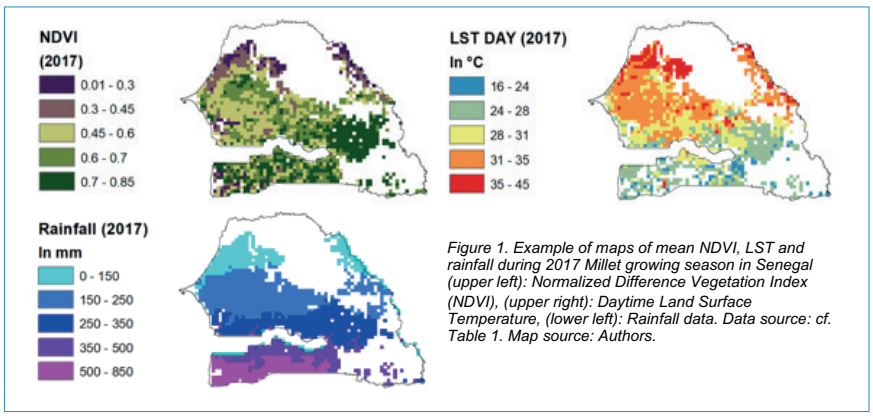


Figure 1. Example of maps of mean NDVI, LST and rainfall during 2017 Millet growing season in Senegal (upper left): Normalized Difference Vegetation Index (NDVI), (upper right): Daytime Land Surface Temperature, (lower left): Rainfall data. Data source: cf. Table 1. Map source: Authors.

One of the most known and used parameters to characterize vegetation covers and thus track crop growth is the Normalized Difference Vegetation Index (NDVI) which is derived from near infrared and red bands of satellites' multispectral (MS) sensors. The NDVI is extensively used due to its close relationship with several vegetation parameters such as Leaf Area Index (LAI), fraction of Absorbed Photosynthetically Active Radiation (fAPAR) and green biomass. The rationale behind NDVI is that crops' leaves absorb red spectrum of visible light as energy source for photosynthesis processes and reflect the infrared spectrum. Therefore, normalized reflectance difference between red and infrared spectrum can assess how healthy a vegetation cover is with the amount of absorbed red and reflected near infrared lights. Many studies have predicted crop yield based on NDVI signals (Rembold *et al.*, 2013), Millet yield assessment in Burkina Faso (Rasmussen, 1992) and Millet production forecast in Senegal (Rasmussen, 1997).

Relying only on NDVI as a proxy for crop yield estimation means that other signals are not exploited for the purpose of food crop production forecasts. Our approach emphasizes the use of several remote sensing products and third-party maps for the same purpose of production forecasting. Several studies conducted during the 1970's have shown that final crop yield can be related to thermal indices (Leroux *et al.*, 2018). We include a daytime Land Surface Temperature (LST) layer as input in the model, in addition to rainfall data retrieved from Climate Hazards group InfraRed Precipitation with Station data (CHIRPS) remote sensing products. Finally, the model uses harvested areas and production maps from the MapSpam database[1], which also includes physical areas, yield, and production value maps for more than 40 crops at global scale.

We apply the model to predict millet production in Senegal for the 2020 season. The three sets of maps present the main inputs used. The first set (Figures 1) shows the mean values of NDVI, LST and rainfall during the 2017

1 https://www.mapspam.info/

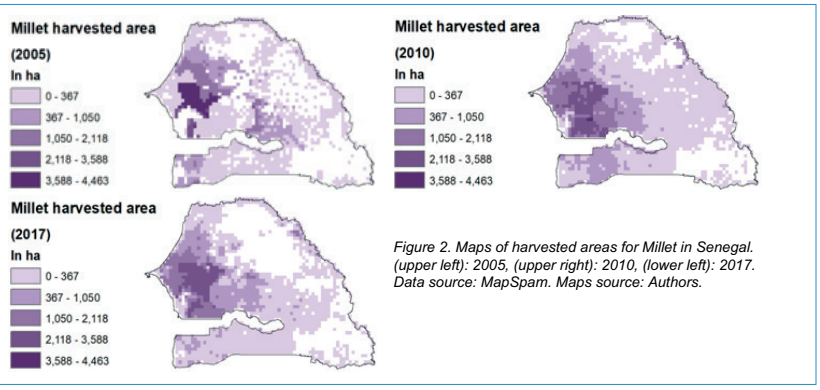


Figure 2. Maps of harvested areas for Millet in Senegal. (upper left): 2005, (upper right): 2010, (lower left): 2017. Data source: MapSpam. Maps source: Authors.

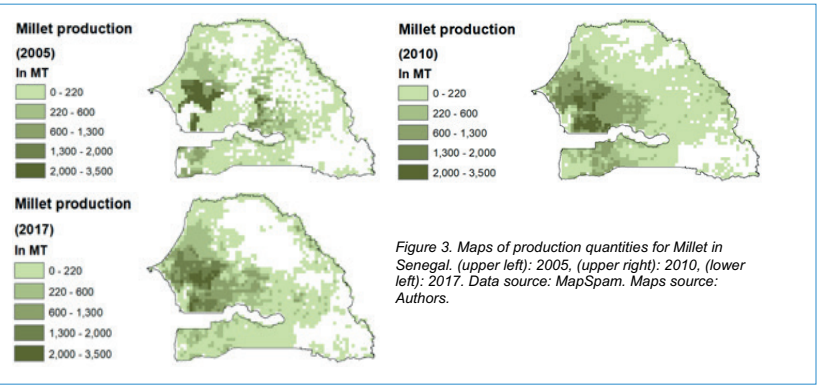


Figure 3. Maps of production quantities for Millet in Senegal. (upper left): 2005, (upper right): 2010, (lower left): 2017. Data source: MapSpam. Maps source: Authors.

Millet growing season. The second set shows harvested millet areas in 2005, 2010 and 2017 (Figure 2). Each pixel contains a value that represents the sum of Millet harvested areas of agricultural lands underneath. The final set (Figure 3) presents the distribution of millet production for the same years. Each map presents pixel values of Millet production for the corresponding year. Harvested areas and production maps have the same pixel locations but with different values and can be used as crop masks to extract explanatory variables at the exact same location where we know with a good probability that Millet has been grown. The model takes NDVI, LST, rainfall and harvested areas as inputs to predict production values. The data used and their sources and main technical characteristics are described in Table 1.

Table 1. *List of our model's features with their source or remote sensing product ID, temporal and spatial resolutions*

Model features	Data source / RS product	Temporal res.	Spatial res.	Type
Normalized Difference Vegetation Index	MODIS[1] (MOD13A2.006)	16 days	1 km	Input
Daytime Land Surface Temperature	MODIS (MOD11A2.006)	8 days	1 km	Input
Rainfall	CHIRPS	Monthly	0.05°	Input
Harvested Areas	MapSpam	Annual	10 km	Input
Production	MapSpam	Annual	10 km	Output

Machine learning and predictive modeling for forecasting under shocks

Machine Learning is a sub-domain of Artificial Intelligence whose recent adoption in several research fields is a consequence of improvements in computations capacities, algorithms and the number of available datasets. Machine learning techniques are designed to mine datasets with the goal of extracting as much information as possible to "teach" the computer programs to recognize and mimic its patterns without being explicitly programmed.

Teaching a computer to learn patterns within datasets involves three main 'learning' techniques that are widely used: supervised, unsupervised and reinforcement learning. The type of technique to be used depends on the type of available data and problem. When explanatory and response variables are available to build a predictive model, supervised learning will be used to build the relationship between inputs and outputs. When only explanatory variables are available, no relationship can be built. Instead, a structural pattern can be found by grouping dataset features into clusters. Such procedure uses unsupervised learning to investigate intrinsic similarities between data points with the mean of measuring how far they are from each other. Reinforcement learning is a sort of combination between the two aforementioned techniques in a way that the response variable values are generated along a process of trial and error. The model generates successive rounds of predictions and regularly updates its parameters to improve its performance in terms of producing the desired responses. For our current purpose, we have used a supervised learning algorithm to predict production as a response variable using NDVI, daytime LST, rainfall and harvested area as explanatory variables.

Data pre-processing

The overarching goal of the data pre-processing procedure is to build the final dataset which will be used to train the algorithm to learn the relationship between features such as NDVI, daytime LST, Rainfall and harvested areas, with the targeted crop production values. The data used are 2005, 2010 and 2017, years for which crop masks are available on the MapSpam portal. The entire set of data-processing steps described below was completed with Spyder-python 3.7.0 provided in the open-source individual Anaconda distribution.

Mosaicking process

The first step of the data pre-processing stage is the mosaicking process which entails merging together different tiles of the same sensing date to cover a specific region of interest. Such a process is specific to satellite images that are sensed with a specific order due to satellite trajectory around the earth. MODIS global sinusoidal tile grid is composed of 595 tiles with 460 that are non-filled. For a region like Senegal, one tile is sufficient to cover the whole country, therefore no mosaicking process is needed.

Raster extraction and cleaning process

The objective of this step is to extract only needed Scientific DataSets (SDS) layers from satellite images and to take out unreliable pixels. For NDVI, layers 1 and 12 have been used to extract NDVI rasters and keep pixels that are labelled as good or marginal data. The same process applies to daytime LST, while rainfall and harvested areas are ready-to-use rasters.

Reprojection, pixel resampling and cropping

At this stage, the methodology consists of performing 3 main operations: reprojection, pixel resampling and cropping. MODIS products that have been selected for the production model are sinusoidal-projected. For further operations with country administrative borders, both shapefiles and rasters need to have the same projection system and Geospatial Data Abstraction Library (GDAL) package has been used to transform each raster projection system from sinusoidal to World Geodetic System 1984 (WGS84) which is the one used for the Global Positioning System (GPS).

In addition, pixel size has to be the same between rasters and crop masks for future operations. MapSpam pixel size has been used as reference to resample input rasters pixels with GDAL resampling procedure. Finally, level 0 shapefile (national level) has been applied to isolate the area of interest.

Crop mask application

Production and harvested areas maps retrieved from MapSpam portal are rasters with the same pixel number location and size. The main difference is in the value they contain. To further isolate explanatory variables at areas where a specific crop is grown, a crop mask has been built using one of the aforementioned one. Such operation consists of replacing *no data* pixels' value with 0, and not null pixels' value with unity. Therefore, by performing the arithmetic product of such a mask with all the maps that have been generated, the result would be new NDVI, daytime LST and rainfall rasters at pixels where the selected crop is located. However, for the 2020 dataset we used the 2017 mask which is the latest available and harvested areas maps do not need to go through such operations since they already have the desired pixel locations.

Generate mean rasters for each feature

To build an agricultural production estimator in a supervised learning manner, explanatory and their corresponding response variables are mandatory. In our case, each line (equivalent to a specific pixel) of the final dataset on which the model will be built on, is a scenario. Therefore, the temporal resolution between inputs and outputs must match. However, production values are available on an annual basis which means pixel values for explanatory variables need to be annualized, and for that mean values were computed for each feature during the growing season. For one crop, the final outputs would be 15 mean (or annual) rasters that are cropped to the region of interest and correspond to 3 rasters (2005, 2010 and 2017) x 5 variables (4 explanatory and 1 response).

From raster to data frame

The final step for the data pre-processing procedure has been to transform generated rasters from the previous section to a data frame that will be used for modeling purposes.

Predicting inputs variables

When building the production estimator, only the mean values of inputs during the growing season for the selected crop are taken into account. Such a procedure is explained first by the need to match the response variable temporal resolution which is annual; and second to only encounter NDVI values that are relevant for crop growth. However, inputs' mean values are not available at the onset of the growing season for the current year (2020) for which we would like to make predictions. Hence, we predicted the various inputs by using a Random Forest (RF) regressor built on top of each inputs' historical data for the last 20 years. Figure 4.a. shows an example of predicted versus actual data on the test set that has been isolated for accuracy assessment for NDVI and, Figure 4.b. shows predicted mean NDVI value for 2020. The choice of random forest regressor is based on its ability to retrieve confidence intervals of forecasts through bootstrap processes which will help in creating bounds and assessing accuracy. Forecasted outputs are then used as inputs for the production model.

Predictive model architecture

For our model we built an Artificial Neural Network (ANN) which is a supervised learning technique. ANNs are inspired by the common representation of human brain neuronal connections. It is a network of unit elements called neurons or perceptrons that perform each specific task. A network consists of an input layer that hosts all input data, an output layer that renders the result, and one or more hidden layers that process the information. Each layer consists of one or more

Figure 4a. *Predicted and Observed mean NDVI values on test set for Senegal and Millet[2].*

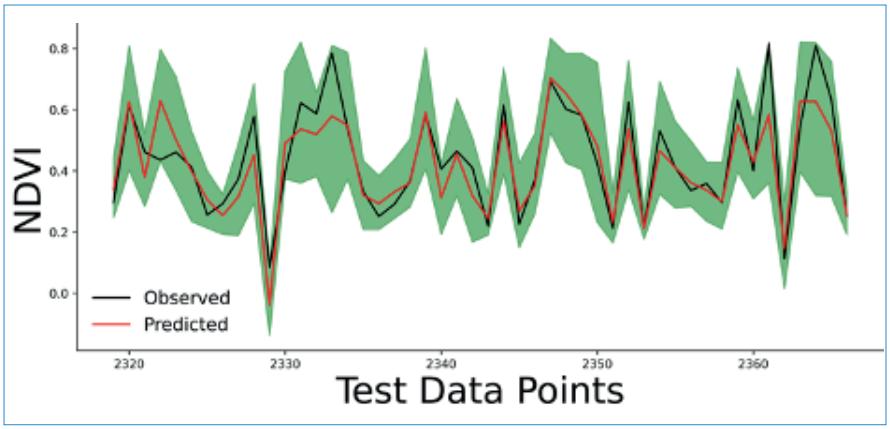
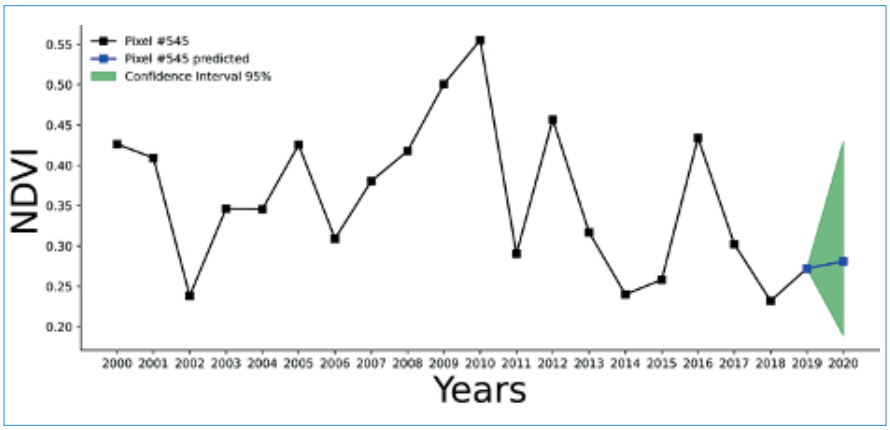


Figure 4b. *Historical and predicted mean NDVI values for Senegal and Millet[3].*



neurons and each neuron processes the information that comes to it through a function called activation function. Each neuron communicates with all the neurons of the previous layer and with those of the neurons of the next layer if it exists. The connections are made possible by weights that are assigned to each bound and updated during the learning phase. Each neuron layer is supplemented with a so-called *biased* neuron, which purpose is to simplify the algorithmic writing. For our production model we used a four-layer neural network: An input layer with 4 neurons that correspond to the 4 input parameters - Two hidden layers with 50 neurons each - An output layer with 1 neuron to forecast one-pixel production value.

The learning process is encountered with a backpropagation algorithm that proceeds by initializing weights with a random and uniform sampling. Then a feed-forward pass is performed to produce a first output that will be compared with the actual outputs. A loss function is computed, and a backward pass is initiated to generate loss gradients with respect to each weight. Therefore, weights are updated using a gradient descent rule and the process is repeated until the exit criteria is reached.

For our production model, we used the mean squared error loss function during the training process along with the Adaptive Moment Estimation (ADAM) optimizer (Kingma & Ba, 2014). For accuracy assessment, we evaluate the model on a test set that corresponds to 10% of the global dataset. A root mean squared error of 0.080 has been obtained for our current example of Senegal and Millet.

4. Production forecasting map: Example of Millet in Senegal

The Covid-19 pandemic makes ground-truth data gathering - to assess food crop production - a difficult task given the restrictions and potential health risks. In addition, the way countries have been generating food production forecasts can lead to biased outcomes due its limitations in disaggregation levels and in their ability to encounter new agricultural lands. Our approach helps to overcome the aforementioned barriers by harnessing remotely sensed data and third-party maps as well as machine learning techniques.

In periods of crisis such as Covid-19 which ensconced several uncertainties about its impacts across sectors in particular the agricultural one, assessing where production basins are is a valuable asset. Countries have long been generating such information at a national level, but for "location-sensitive" crisis such as Covid-19, disaggregation is a key element in mitigating its effects on communities. Therefore, producing such a map of food crop production has the potential to make Covid-19 agriculture-oriented policy responses more community-oriented. Its value is even more perceptible when production is predicted prior to crop harvesting periods.

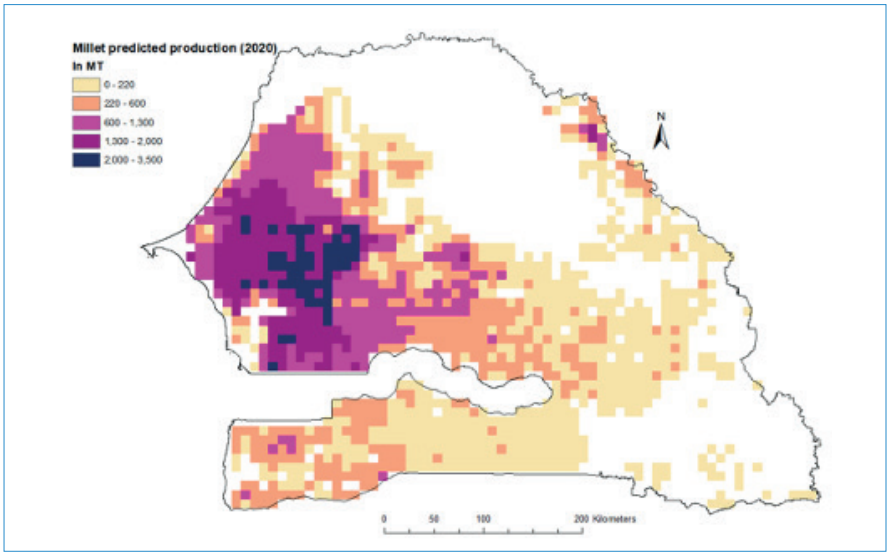
Input data for the current year (2020) have been produced using the RF model. They correspond to estimated mean NDVI, daytime LST and rainfall data during the growing season for Millet in Senegal. Such values have been fed to the model to forecast Millet production along with 2017 harvested area data. Their pixel IDs have been used as dummy variables in order to reassemble the predicted values into a new raster.

The map illustrated on Figure 5 is our model's output and represents the Republic of Senegal with pixels of Millet production for the current season. Each pixel embeds a predicted production value that corresponds to the amount of Millet that is expected to be grown underneath.

Conclusion

The model we presented here uses remote sensing products and machine learning techniques to predict food production for a specific crop and country. While

Figure 5. *National Millet production (Senegal) forecasted with our model for the current 2020 season. Production units are in metric tons, and pixels are of size 10km x 10km.*



food crop production has long been estimated at national or administrative borders' level, the main advantage provided with the aforementioned model is its ability to forecast production at pixel level. Such open new possibilities in food security assessment, food balance sheet and in designing targeted policies, specifically during the ongoing COVID-19 pandemic.

An example of Senegal and for Millet were taken to succinctly illustrate the methodology. In the coming weeks and months, more countries will be covered across the continent and the staple food crop production forecasts will be made available in future bulletins and newsletters under this particular workstream.

References

Huang, J., & Han, D., (2014). Meta-analysis of influential factors on crop yield estimation by remote sensing. *International Journal of Remote Sensing*, 35(6), 2267-2295. doi:10.1080/014311612014890761

Rembold, F., Atzberger, C., Savin, I., & Rojas, O. (2013). Using Low Resolution Satellite Imagery for Yield Prediction and Yield Anomaly Detection. *Remote Sensing*, 5, 1704-1733. doi:10.3390/rs5041704

Rasmussen, M. S. (1992). Assessment of Millet Yields and Production in Northern Burkina-Faso Using Integrated NDVI from the AVHRR. *International Journal of Remote Sensing*, 13 (18), 3431–3442. doi:10.1080/01431169208904132.

Rasmussen, M. S. (1997). Operational Yield Forecast Using AVHRR NDVI Data: Reduction of Environmental and Inter-Annual Variability. *International Journal of Remote Sensing*, 18 (5). 1059–1077. doi:10.1080/014311697218575.

Leroux, L., Baron, C., Zoungrana, B., Traore, S., Lo Seen, D., Lo Seen., & Begue, A. (2018). Crop Monitoring Using Vegetation and Thermal Indices for Yield Estimates: Case Study of a Rainfed Cereal in Semi-Arid West Africa. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(1). 347-362. Cirad-01951499.

Kingma, Diederik & Ba, Jimmy. (2014). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations.

Figures notes

1 Moderate Resolution Imaging Spectroradiometer (MODIS)

2 White pixels correspond to areas where Millet is not grown or nodata values that have been removed from remote sensing products.

3 Green areas correspond to 95% empirical confidence intervals.

The Expertise we need

AKADEMIYA

The Africa we want

Recommended citation: Racine Ly, Khadim Dia. 2020. Application of Remote Sensing and Machine Learning for Crop Production Forecasting During Crises. Covid-19 Bulletin No. 4, August. Kigali. AKADEMIYA2063.

Note: *The boundaries and names shown, and the designations used on maps do not imply official endorsement or acceptance by AKADEMIYA2063.*

AKADEMIYA2063 is grateful to USAID for funding for this work through a Feed the Future grant with Policy LINK. Any opinions stated here are those of the author(s) and are not necessarily representative of or endorsed by AKADEMIYA2063.

a: AKADEMIYA2063 | Kicukiro/Niboye KK 360 St 8 | 4729 Kigali-Rwanda
p: +221 77 761 73 02 | p: +250 788 304 270 | e: hq-office@akademiya2063.org | w: akademiya2063.org