Vol.1 Issue-12, AUG 2021 (e-ISSN: 2582-8223)

JUST AGRICULTURE
multidisciplinary e-Newsletter

Critical Factors Contributing to the Spread of Fall Armyworms from a Data Science Perspective

Dr. Shravani Basu, Ángel de Jaén Gotarredona, Dr. Sébastien Foucaud & Dr. Mukti Sadhan Basu. SBSF Consultancy

ARTICLE ID: 019

Goals of the study:

Although very limited information is available about the outbreaks of Fall Armyworm (Spodoptera frugiperda; FAW), which is fast becoming ineradicable across continents, the only way then to mitigate the monumental losses left in the wake of its infestation is to understand the pest in great details and develop multi-pronged crop protection and pest control strategies. This study attempts to shed light on the conditions favouring the spread of FAWs by combining three different datasets bringing complementary information on potential outbreaks of FAW.

Due to the limitations of the data collection method used by FAO, the combined dataset does not allow to build a forecasting model. This study, therefore, focuses on extracting features which are driving the spread of FAW in Africa, where most of the data has been collected. Although based on the Africa dataset, we have conducted the analysis in such a way that the findings can be extended to any FAW incidences globally.

Fall Armyworms: a global threat:

The larval stage of FAW moth is a pest that feeds on more than 350 plant species, causing extensive damage to economically important cultivated cereals such as maize, rice, sorghum; ash crops like cotton, sugarcane, peanuts; fruit crops like apples and oranges, and vegetable crops, among others. Maize however remains the preferred host. Because the caterpillar eats so much of the plant, they are very detrimental to crop survival and yield. According to the FAO, as much as 18 million tonnes of maize are lost annually in Africa, enough to feed tens of millions of people for whom it is a staple crop and representing an economic loss of up to

$4.6 billion in the continent. In addition, country specific studies have shown a positive correlation between FAW exposure and intensity of insecticide use._Several measures to control FAW infestation exist and yet do not significantly reduce the losses.

In India, FAW was first reported in the fields of Chikkaballapur district in Karnataka according to a survey conducted by the National Bureau of Agriculture Insect Resources (NBAIR) in July 2018. The pest has destroyed more than 70 percent of the crop in Karnataka and has now spread further into southern, western, northern and north-eastern India. Maize is grown in about 9.3 million hectares with a total annual production close to 28 million tons.

Datasets:

We have combined three complementary, publicly available datasets for this study.

1. The core data set used in this study is the data being collected under the program launched by the Food and Agriculture Organization (FAO) of the United Nations, for **Global Action for Fall Armyworm Control**, which catalogues cases of FAW outbreaks, mostly in African countries. The **FAW Monitoring and Early Warning System (FAMEWS)** consists of a mobile app distributed to farmers for data collection and a global platform for mapping the current situation. The data is mostly collected by farmers themselves directly in the FAMEWS app, using two techniques of detection: collecting insects using pheromone traps or scouting the field. The version of the datasets used for this study cover cases registered between 27-02-2018 and 30-09-2019. The selected dataset consists of 39013 cases and for each case 44 variables have been measured. Unfortunately, this dataset from FAO is no longer freely available on their website anymore.

2. In combination to the FAW outbreak dataset, we used reanalysed Weather Data from the VIC model in the Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS). The data are in 0.25-degree resolution covering the entire African continent from January 2001 to present (in case of this study we used data available between February 2018 and October 2019). The temporal resolution is daily. A total of 21

variables are extracted from the model, comprising precipitation amount, temperature and wind speed.

3. The last dataset used is the Soil Data from Harmonized World Soil Database (HWSD). It is a 30 arc-second raster database with over 15000 different soil mapping units that combines existing regional and national updates of soil information worldwide (SOTER, ESD, Soil Map of China, WISE). The datasets comprise 58 variables, describing the composition in terms of soil units and the characterization of soil parameters (organic Carbon, pH, water storage capacity, soil depth, cation exchange capacity of the soil and the clay fraction, total exchangeable nutrients, lime and gypsum contents, sodium exchange percentage, salinity, textural class and granulometry).

The three datasets were merged based on the geo-coordinates of the crop fields provided in the FAMEWS data.

Method

A. Selection:

Two different methods of inspection have been used in the FAMEWS dataset: Scouting and Pheromone traps. The different inspection methods, as standalone and in combination, may generate unforeseen biases. We have, therefore, limited the study to the sample detected using Scouting, as it represents the largest sample (69% of the cases, i.e., 26901).

As shown in Figure 6, when looking at the relative distribution of positive and negative detections of FAW in the Scouting sample from FAMEWS, positive cases are dominating the sample at 85.9%. The very strong bias for positive detections in the dataset actually suggests that the data collection through the mobile app has been started after the actual FAW outbreaks. If the data collection was made from a systematic survey in a given area (with inputs from all farmers in that area during the survey period, irrespective of the incidence of FAW), we would expect to see a far lesser fraction of detections in the dataset.

As demonstrated, the Scouting-inspection-based dataset is highly biased toward positive detections. This is most likely due to the reactive nature of data collection following outbreaks (farmers only use the app once their field has been infested). Such an unbalanced dataset doesn't allow us to build a prediction model of FAW outbreak for a given field at a given date. To achieve such a goal, a more systematic and unbiased data collection is necessary, to provide a realistic representation of outbreak cases.

We, therefore, decided to focus our research on the drivers and aggravating factors of the spread of FAW in maize crops. In practice, we are using machine-learning modelling to extract a set of features with predictive power to identify the presence of FAW. We want to make sure the predictive importance of these features can be extrapolated through location and time, so that the insights we get from Africa's outbreaks in the past are valid for other countries (e.g., India) in the future.

B. Model, Pre-processing & Validation strategy:

We decided to train an Extreme Gradient Boosting optimized to predict, as a target variable, the percentage of plants infested by FAW at a given inspection point. As described in the previous section, prediction of the spread of the pest (i.e., if a field is going to be affected or not?) is not possible due to the bias in the detection sample. Therefore, we chose to focus on the prediction of the level of infestation (i.e., knowing that the field is infested, what is the expected fraction to be affected?)

Before any model training, we proceed for a proper pre-processing of the dataset:

- Limiting the dataset to the African continent and maize crops (as information on other crops are also present in the dataset) - reducing the total sample to 16705 cases;
- Removal of duplicate rows and constant columns;
- Domain-driven selection of features from the FLDAS, FAMEWS and HWDS datasets (focusing only on relevant information, dropping irrelevant or duplicated features, e.g., database IDs);
- Standardization of certain fields (e.g., 'cornfields' as different measurement units are used depending on the user of the app);

- Manual feature engineering, combining features in a more relevant fashion for our study;
- Aggregation of weather data to reflect weekly averages.

Given the nature of the FAMEWS dataset, and the mixture of spatial and temporal features used (soil and weather data), we have to implement a particular validation strategy, to ensure that the most important features will generalize properly through time and location. Our validation strategy is based on a temporal separation, where data from 2018 is used in training and data from 2019 in validation. However, to decouple any spatial effect, we improve this basic strategy as follows:

- For the validation sets from 2019, we define 3 specific areas as intervals of longitudes: A (-16°, -1°), B (26°, 33°) and C (35.5°, 46°). These areas have been carefully designed to maximize the amount of both training and validation instances.
- We then split our dataset of inspections in training/validation three times, each time training the model on the 2018 dataset excluding data in the validation area (e.g., all 2018 data outside of area A), and validating with the 2019 data set within the validation area (e.g., all 2019 data within A).
- With such a validation strategy, we optimize the most important hyper-parameters starting with a global grid search, followed by a more local grid search to fine-tune. With the three splits described above, while fine-tuning the hyper-parameters of the model, we can calculate the Mean Absolute Error (MAE) of the model between the validation sets. The lower the MAE, the higher is the ability of the model to generalize through location and time.

C. Feature selection:

Once the model is trained, we can calculate, for each feature, the gain in the XGBoost objective function obtained when one of the decision trees of the model makes a split in the dataset using that feature. We then define the importance of a feature within a XGBoost model as the sum of all these gains.

However, feature importance can be artificially boosted by hidden correlations, or masking information about location or time. To ensure that the selected features generalize properly, for each feature, we again fine-tune and train a model leaving the feature out. We then compare the MAE of this new model excluding the feature, with the previous model including it. If the MAE increases considerably (more than 0.005) when leaving the feature out, we discard the feature. In this case, the set of important features is re-calculated again, leaving the discarded feature out of our predictors set.

Analysis & Results:

To visualize the power of the selected features to discriminate between highly and slightly infested crops, we build a decision tree using a binary target which is defined to be True if the crop has a percentage of plants with FAW over the median (which is around 25% of infestation), and False otherwise, as represented in Figure 10.

The decision tree reads from top to bottom and as follows:

- Each cell represents a split into two parts of the sample based on the stated condition (for instance on the top cell 'Psurf_f_tavg_mean<96627.656'); if condition is satisfied (True) then the corresponding splitted sample is checked against the next level cell on the left, if not satisfied (False) on the right;
- In each cell the fractions of negative (in our case less infested) and positive (in our case more infested) are indicated in brackets (for instance on the top cell 'value= [0.513,0.487]' hence 51.3% of negatives and 48.7% of positives);
- The proportion of the full sample represented by each cell is as well indicated (for instance in the top cell 'samples=100%');
- The higher the fraction of positives (more infested) the bluer is the cell, the higher the negatives (less infested) the redder.

Impact of soil density:

Here we need to highlight that some of these features have a real predictive power only when combined with others. For example, as seen in Figure 11, the fraction of Clay 'T_CLAY' independently has a rather weak correlation with our target. However, the second plot in Figure 11 shows that when combined with the surface pressure 'Psurf_f_tavg_mean' it has a stronger impact. This fact can already be deduced by the decision tree.

A tentative interpretation is that at higher Surface atmospheric pressure the soil is denser disregarding its composition. At lower atmospheric pressure, the soil being less dense, then the texture of Clay will impact more the density. With a less dense soil (so at low atmospheric pressure and lower fraction of Clay content), the FAW adult moth can more easily emerge from the soil once the pupae have enclosed, increasing the risk of infestation.A similar interpretation can be put forward with the fraction of Organic Content as seen in Figure 12: the higher the Organic Content fraction the less dense the soil will be, therefore the higher the risk of infestation.

Note that in the figures 11 and 12, we used a third order regression. The low statistics in measurements in the extreme cases of Clay or Organic Content weight fraction, prevent any interpretations as such measurements and may not be reliable enough.

Impact of the stage of crop growth and crop health:

The FAW caterpillars feed mostly on the tender part of the maize leaves and whorls, which explains a stronger risk of infestations on young crops and a rapid decrease of the risk with older crops seen in Figure 13. The peak of infestation is around 30-80 days. Maize matures in 130-135 days after planting, which explains the rapid decrease in statistics (and infestation cases) beyond this age, as such cases are probably mostly related to mistakes in data collection (as the age is computed from data collected by the farmers directly).

In Figure 14, we investigated the impact of Soil temperature on infestation rate (air temperature and the soil radiative temperature correlate very strongly with soil temperature so we focused only on this variable).

Impact of weather conditions:

When displaying the effect of wind on infestation, as shown in Figure 15, an increase of infestation is noticeable at first with stronger winds (up to 2m/s), but when wind is getting stronger the infestations are reducing (up to 4m/s). A potential interpretation would be that some wind improves the chances of the FAW moth to spread and resume the cycle in a different area of the crop field, but stronger winds prevent FAW caterpillars to remain on the leaves to continue feeding.

Figure 16 shows a negative correlation between the infestation and both Air Specific Humidity and Rainfall. These results indicate that despite humidity and rainfall being beneficial for the development of the plant, excessive rain washes out FAW caterpillars from the leaves, in fact reducing the infestation rate. This is a controversial result as some studies in the literature tend to indicate the opposite.

Impact of irrigation:

As seen in Figure 16, Rainfall seems to be slowing down infestation. We further investigate the impact of watering to see if indeed rainfall is harmful for FAW.

In Figure 17-*Left*, Soil moisture displays similar negative correlation with infestation rate as Air humidity and Rainfall. However, when looking at the effect of different types of watering, in Figure 17-*Central*, it is clear that while irrigation seems to favour infestation, Rainfed watering is definitely reducing the infestation rate, confirming the result mentioned above. As shown in Figure 17-*Right*, Soil moisture is on average higher with Rainfed than with Irrigation, which means that rain is favouring the development of the plants as well (in fact soil moisture also depends of soil composition, information provided by the Available Water Capacity 'AWC_CLASS' in our dataset.)

One clear observation which can be made regarding irrigation from this analysis is that any irrigation system (like sprinkler) mimicking rain would reproduce the effects noticed with Rainfed (reducing infestation rate).

It is also interesting to note that there is no evidence from our analysis that lack of soil moisture (water stress) aggravates the infestation by FAW, as reported from South Sudan in the FAO report referenced above. More data and deeper investigations would be required to understand such a discrepancy.

Conclusions and Recommendations:

The work conducted, besides providing some actionable insights (for instance, with irrigation), demonstrates the significance of taking a data science-based approach to use various sources of information, beyond the scope of restricted surveys, to support the development of comprehensive and result oriented agricultural projects.

1. We highlight the importance of a holistic approach by combining distinct but highly complementary datasets (input from farmers "FAMEWS", weather "FLDAS" and soil data "HWSD"), in deriving a consistent and robust picture. As seen in the study, all important features are extracted from the three datasets, and most insights are provided based on a combination of features from the various data sources (as can be seen from the Decision Tree).

2. This entire work is based on the availability of Open Access Data. We would like to acknowledge the work of the teams who built these datasets (data collection, data analysis, modulization and simulations). We want to re-emphasize the importance of sharing such data freely, and are saddened by the recent removal of access to the FAMEWS dataset by FAO from their website (as per the latest status, which up to end 2019 was still possible to download in csv/excel format).

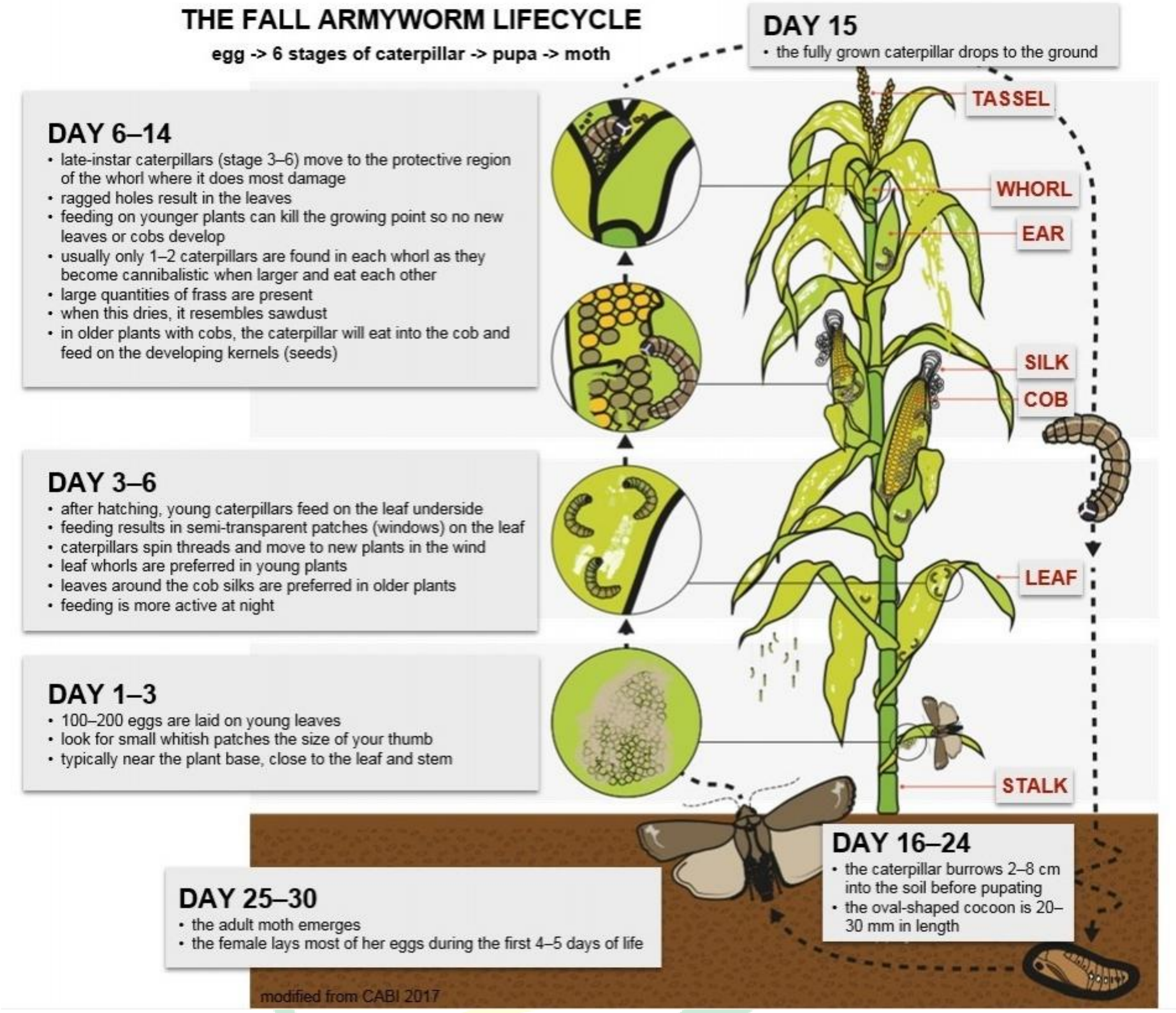
3. We have identified a defined set of important features (14 features), which can be used to better understand the drivers behind the spread of FAW. Although we recommend aggregating as much data as possible, these features can be used as guidelines for future data collection strategies. Most of the results of the analysis conducted here are of course backed by previous studies and common knowledge, but this analysis provides quantifiable information which can be used for building predictive models and identifying actionable measures to limit the spread of the pest.

4. Data acquisition is a critical element of any research, and the excellent work done by FAMEWS deserve to be emphasized here. Providing farmers access to an app to inform them of the status of their field is a major weapon against the spread of pests such as FAW. However, the dataset presents its own limitations due to "informal" key-in by farmers, leading to certain lack of accuracy and generating uncontrollable biases. The collection strategy taken by FAMEWS implied an important bias for post outbreak measurements, as

farmers used the app mostly after an incidence of infestation, which prevents the possibility to build a predictive model of the FAW spread.

We would recommend adopting a systematic survey strategy by collecting information directly at the farm level, independently of pest infestations over several seasons, and combining the inputs from farmers with independent weather and soil information. The design of the survey needs to have predictive and prescriptive modelling as an aim based on which measures can be adopted to fight biases as much as possible. Such projects require the involvement of Data Scientists and Machine-Learning experts right from conception.

JUST AGRICULTURE
multidisciplinary e-Newsletter



THE FALL ARMYWORM LIFECYCLE

egg -> 6 stages of caterpillar -> pupa -> moth

DAY 15
- the fully grown caterpillar drops to the ground

DAY 6–14
- late-instar caterpillars (stage 3–6) move to the protective region of the whorl where it does most damage
- ragged holes result in the leaves
- feeding on younger plants can kill the growing point so no new leaves or cobs develop
- usually only 1–2 caterpillars are found in each whorl as they become cannibalistic when larger and eat each other
- large quantities of frass are present
- when this dries, it resembles sawdust
- in older plants with cobs, the caterpillar will eat into the cob and feed on the developing kernels (seeds)

DAY 3–6
- after hatching, young caterpillars feed on the leaf underside
- feeding results in semi-transparent patches (windows) on the leaf
- caterpillars spin threads and move to new plants in the wind
- leaf whorls are preferred in young plants
- leaves around the cob silks are preferred in older plants
- feeding is more active at night

DAY 1–3
- 100–200 eggs are laid on young leaves
- look for small whitish patches the size of your thumb
- typically near the plant base, close to the leaf and stem

DAY 16–24
- the caterpillar burrows 2–8 cm into the soil before pupating
- the oval-shaped cocoon is 20–30 mm in length

DAY 25–30
- the adult moth emerges
- the female lays most of her eggs during the first 4–5 days of life

modified from CABI 2017