Machine Learning-based Locust Swarm Prediction in Ethiopia using Landsat and FLDAS Climate Dataset

Ka Hei Chow[1] (ka-hei-pinky.chow@stud-mail.uni-wuerzburg.de), Narges Mohammadi Khoshouei[2] (narges.mohammadi_khoshouei@stud-mail.uni-wuerzburg.de)
Department of Remote Sensing, Institute of Geography and Geology, Julius-Maximilians-Universität Würzburg (JMU) [1,2]

Abstract

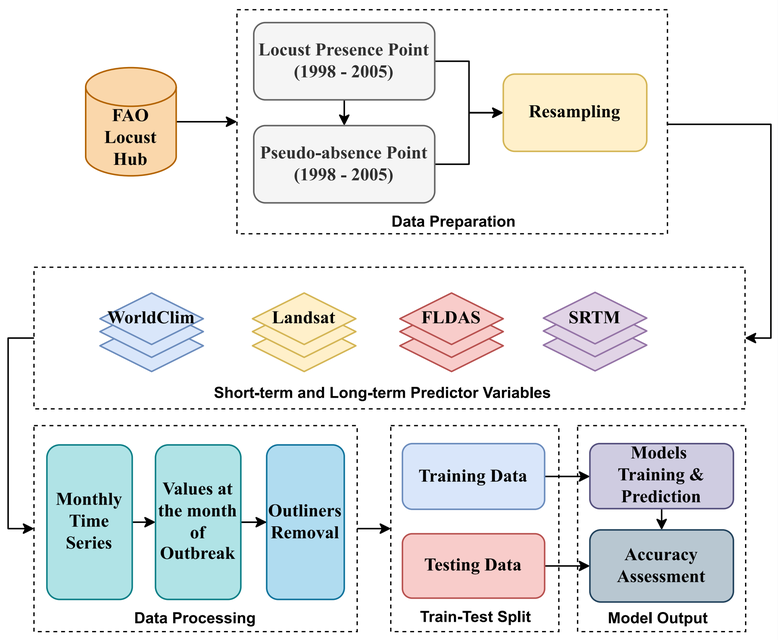
Desert locust (Schistocerca gregaria sp.) invasion has frequently caused huge economic loss and threatened food security and population livelihood in the North African countries (Gómez, D. et. al, 2019). The current literature largely focuses on ecological niche modelling and monitoring locust plague with temporal dynamics of vegetation indices, limiting the prediction performance due to the insufficient considerations of land use and species' favourable conditions (Klein, I., Oppelt, N., & Kuenzer, C., 2021). As a result, relevant factors such as land use, phenology and seasonal variations which play important roles in locust behavioural ecology cannot be well represented. This study aims to bridge this gap by integrating long-term and short-term hydroclimatic variables (FEWS NET Land Information System and WorldClim) and multispectral reflectance (Landsat) into machine learning models, aiming to predict locust plague in Ethiopia. Multiple models are trained using locust data from FAO Locust Hub and the derived variables, yielding optimal accuracy 88.4% with the use of random forest algorithm. The present study shows that model-derived short-term hydroclimatic variables could be effectively combined for locust management and swarm prediction is a relevant use case for FLDAS dataset.

Introduction

For many developing countries, the desert locust has been a serious crop pest. The swarms migrate extensively and over very long distances, threatening the food security of huge populations. For instance, the locust plague in 2004 had reduced over 80% of the crop production in Burkina Faso, Mali, and Mauritania, leading to the acute need for external aid. Hence, developing an accurate early warning system is critical and can largely benefit from the open earth observation (EO) data. The existing research regarding EO-based locust management is, nevertheless, limited and highly biased geographically (Klein, I., Oppelt, N., & Kuenzer, C., 2021). Most of the research focuses on land use information, long term bio-climatic data and vegetation indices (eg. EVI, NDVI), which restricts the uses of multi-spectral information as well as short-term climatic data (Ceccato, P., 2005, Waldner, F. et. al., 2015). In the present study, data between 1998 and 2005 across Africa, together with Landsat and climatic variables, are used for machine learning model training before applying the models in Ethiopia for swarm prediction and accuracy assessment. In this way, the influences of short-term climatic conditions, such as soil moisture dynamics can be better considered for the locust behavioural ecology.

Fig 1. Locust occurrence and Overview Map

Workflow



Methodology

Data Acquisition and Pre-processing

This study uses swarm data from Locust Hub, published by the UN FAO in 2020 and includes global coordinates over the last 4 decades. The data between 1998 and 2005 in Africa is pre-processed in Python, including water masking and data argumentation with random jitters at the individual point level. Pseudo-absence data is generated by resampling non-presence data for each month and year, with the assumption that swarms are absent given no record within 1,000 kilometres for the entire year. The prepared data consists of approximately 32,000 and 37,000 points for the presence and absence classes.

Climate Dataset: WorldClim and FLDAS

WorldClim provides bio-climatic variables layers aggregated across the last few decades, while FLDAS, the Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System, is a model-driven dataset combining both reanalysis, satellite and in-situ data from MERRA-2 and CHIRPS. It aims to estimate hydro-climate conditions specifically for sub-Saharan Africa. Selected variables including soil moisture, the temperature at different soil depths, as well as surface air temperature and wind speed, are used to train multiple machine learning models. Other training data includes spectral data from Landsat 5,7, and 8, as well as SRTM DEM.

Google Earth Engine (GEE)

Climate variables over 7 years for the data coordinates are extracted in the cloud. The data coordinates are first imported into Google Earth Engine (GEE). Then, Landsat bands and FLDAS variables are exported over the whole study period as monthly time series at individual coordinates. Meanwhile, WorldClim data is downloaded from the web before point extraction. All variables apart from DEM are further filtered to the month of outbreak. All variables are joined in a tabular format before outliner removal and train-test split at a ratio of 70:30.

Machine Learning Model

After data preparation and cleaning, multiple supervised machine learning algorithms, including Naive Bayes, Support-vector machine, Logistic Regression, KNN, Decision Tree and Random Forest, are selected for the high-dimensional binary classification problem. The classifiers are built using scikit-learn Python library on Google Colab. The models are trained with optimized parameters for binary predictions using the test data in Ethiopia.

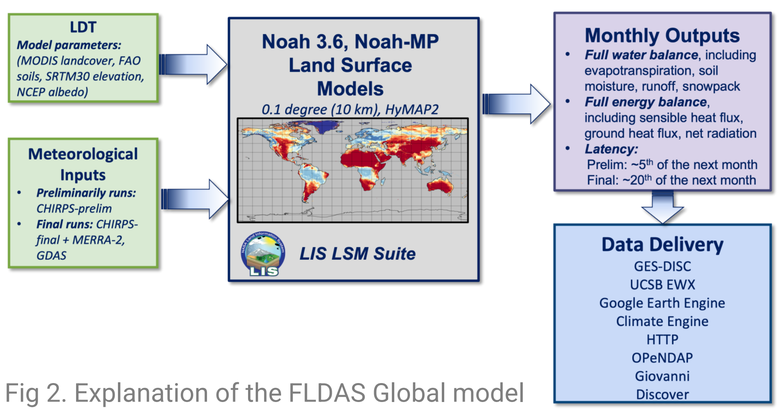


Fig 2. Explanation of the FLDAS Global model

Predictor Variables	
Landsat RGB Bands	Evapotranspiration
NIR Band	Surface pressure
SWIR Band	Specific humidity
SWIR2 Band	Soil heat flux
SRTM DEM	Sensible heat net flux
Max Temperature of Warmest Month	Latent heat net flux
Min Temperature of Coldest Month	Total precipitation rate
Soil moisture (0 - 10 cm/ 10 – 40 cm/ 40 – 100 cm/ 100 – 200 cm underground)	Soil temperature (0 - 10 cm/ 10 – 40 cm/ 40 – 100 cm/ 100 – 200 cm underground)
Near surface air temperature	Near surface wind speed

Results

Multiple machine learning models yield accuracies between 60.4% to 88.4%, with random forest and KNN algorithms performing the best. Overfitting tendencies are observed which would be potentially explained by the specific climatic conditions in the individual years. Besides, the incomplete representation of climatic conditions in the absence data and the class bias (presence as a minority class) added challenges to the predictive models.

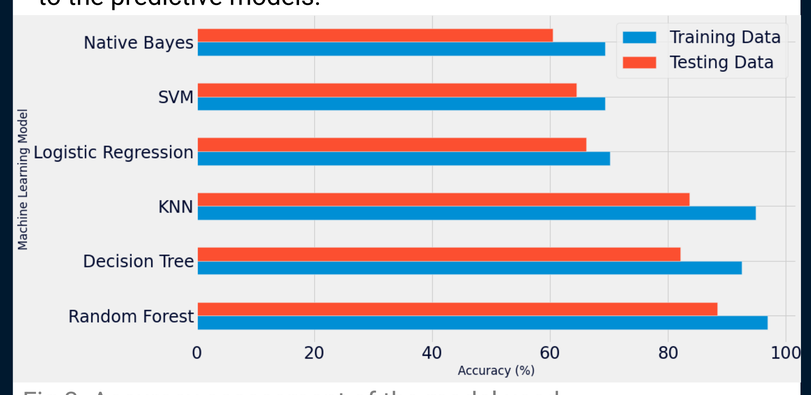


Fig 3. Accuracy assessment of the model used

Conclusion

Improving the swarm predictive model shall take account of seasonal dynamics which shape species' behaviours. This study aims to integrate Landsat and FLDAS, a set of model-driven hydroclimatic variables to predict locust swarms in Ethiopia. It is illustrated that locust management, beyond drought and reservoir monitoring, is a highly relevant use case for the FLDAS data. The developed workflow has also shown clear potential in extracting short-term climatic patterns for machine learning applications using cloud computing. Machine learning models yield accuracies between 60.4% to 88.4%, with the random forest algorithm serving as the optimal classifier, followed by the KNN classifier.

The extracted data points using GEE would be potentially extended to longer time series, as well as to consider climatic conditions in a sequence. Besides, further bio-climatic variables, such as MODIS NDVI and satellite-derived LAI, can be integrated into the model. Besides, it is suggested to consider multi-temporal hydroclimatic variables in the model to better capture phenology and seasonal variations. Further efforts would also utilize deep learning-based models to extract abstract features from spectral reflectance.

Acknowledgements

We thank Dr. Martin Wegmann (Assistant Professor at the Department of Remote Sensing, EAGLE M.Sc. Coordinator) and Schwalb-Willmann, Jakob, M.Sc. (Lecturer for course Advanced Programming of Remote Sensing in which this project was initiated) for their support.

GitHub Repository

https://github.com/pinkychow1010/DeepLearning_LocustPrediction

References

Ceccato, P. Operational Early Warning System Using Spot- Vegetation and Terra- Modis To Predict Desert Locust Outbreaks. In Proceedings of the 2nd International VEGETATION User Conference, Antwerp, Belgium, 24–26 March 2005; pp. 33–41.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International journal of climatology, 37(12), 4302-4315.

Kimathi, E., Tonnang, H. E., Subramanian, S., Cressman, K., Abdel-Rahman, E. M., Tesfayohannes, M., ... & Kelemu, S. (2020). Prediction of breeding regions for the desert locust Schistocerca gregaria in East Africa. Scientific Reports, 10(1), 1-10.

Klein, I., Oppelt, N., & Kuenzer, C. (2021). Application of remote sensing data for locust research and management—A review. Insects, 12(3), 233.

McNally, A., Arsenault, K., Kumar, S., Shukla, S., Peterson, P., Wang, S., ... & Verdin, J. P. (2017). A land data assimilation system for sub-Saharan Africa food and water security applications. Scientific data, 4(1), 1-19.

Waldner, F.; Babah Ebbe, M.A.; Cressman, K.; Defourny, P. Operational Monitoring of the Desert Locust Habitat with Earth Observation: An Assessment. ISPRS Int. J. Geo-Inf. 2015, 4, 2379–2400