Prediction of locust swarms using Machine Learning

By: Sunaina and Bibhu

Locust Swarming and its Environmental Impact

- Locust swarming is a behavioural phase transition problem in ecology
- Population can shift between alternative stable states depending on density
- Swarming and recession are the two stable states of locust populations
- Locust swarms can decimate crops and pastures in a short amount of time
- This leads to famines in developing countries and affects local livelihoods

Implementing Machine Learning Models to Predict Locust Swarms

- Implementing baseline models to understand locust swarming
- Different environmental variables impact locust swarming
- Extrapolating the model to Latin America, India, and other countries with a gap in predicting locust swarms
- Our work provides insights into the ecology of locust swarms
- Generalisability of machine learning models can help predict locust swarms

Papers	Countries used	Features used
Prediction of breeding regions for the Desert Locust Schistocerca Gregaria in East Africa.	Morocco, Mauritania and Saudi Arabia for training and Kenya and Sudan for testing	Temperature, rainfall, soil moisture, and sand content for prediction of Hoppers.
Prediction of desert locust breeding areas using machine learning methods and smos (MIR_SMNRT2) near real time product.	30 countries	Soil moisture for prediction of nymph population
Modelling Desert Locust presences using 32-year soil moisture data on a large-scale	30 countries	Soil moisture for prediction of nymph population.
Machine learning approach to locate desert locust breeding areas based on ESA CCI soil moisture	Mauritania	Soil moisture for prediction of nymph population.
On pseudo-absence generation and machine learning for locust breeding ground prediction in Africa	East African countries	Soil moisture (at different depths), average temperature, wind, rainfall and quality of air.*

Methodology followed - Pre-processing and Feature Engineering

- Time series data from 1985-2021 collected from FAO's locust swarming dataset
- Data comprises hopper absence/presence at global coordinates via Desert Locust Information Service
- 95 days of environmental data prior to hopper presence is scraped
- Statistical descriptions used to engineer new features (mean, median, max, min)
- Time intervals (6, 12, 16, etc.) created from -95 to 0 days
- Environmental variables (temperature, precipitation, soil moisture) scraped from meteorological satellite datasets
- Features undergo suitable pre-processing (centering, scaling)
- Model trained on one set of countries and tested on another set of countries

Pseudo-generation of absence points

- It's difficult to ascertain the absence of a species in an area during ecological surveys
- Researchers generate absence points near presence zones using random sampling or environmental profiling
- Absence points are important for feeding datasets in machine learning models to avoid over-representation of one class
- Some machine learning models use presence-only data, like the MaxEnt species distribution model
- MaxEnt generates "background" points, but doesn't associate them with the absence of the species
- MaxEnt aims to map optimal environmental parameters with the presence of the species.

Different Models and Their Results

- Logistic regression, k-Nearest Neighbors, MaxEnt, XG-Boost are some of the machine learning models used in the literature, The performance of these models varies depending on the countries and features used for testing.
- Soil moisture was found to be a good predictor even when used without any other variable. This table presents the accuracy scores obtained from various models.

Statistic	Logistic	k-NN	Random Forest	MaxEnt
Accuracy	0.85	0.81	0.78	0.81

- Logistic regression and RF are implemented as baseline algorithms.

Rows	Features	Temporal	Non-temporal
31251	1168	Average temperature, wind speed, soil moisture, precipitation, Air humidity	Sand content

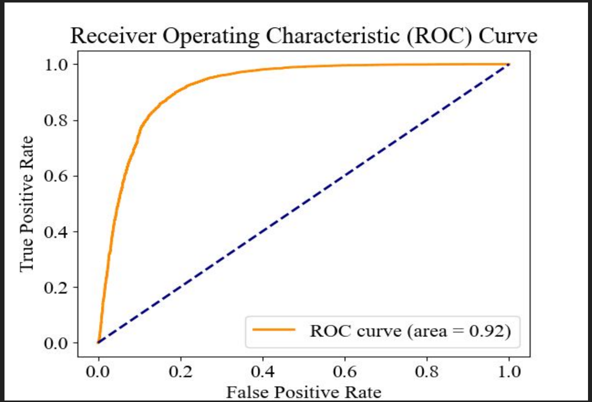
Curation and preprocessing of Dataset

- Pre-processing pipeline from previous works was used to curate the dataset of African countries from FAO's hopper observation data.
- X and Y coordinates were used to fetch data from GLDAS Noah Land Surface Model and SoilGrids for 95 days prior to the presence data.
- Further bucketizing based on a time interval of 6 days created a total of 1168 features.
- The dataset was split into two subsets for training and testing, with a test size of 0.20.
- Two baselines: logistic regression and random forest
- Features used:
    a. Temperature, precipitation, soil moisture, and other environmental variables
    b. Statistical descriptions of the features (mean, median, maximum, minimum)
- Default L2 penalty term used in logistic regression.

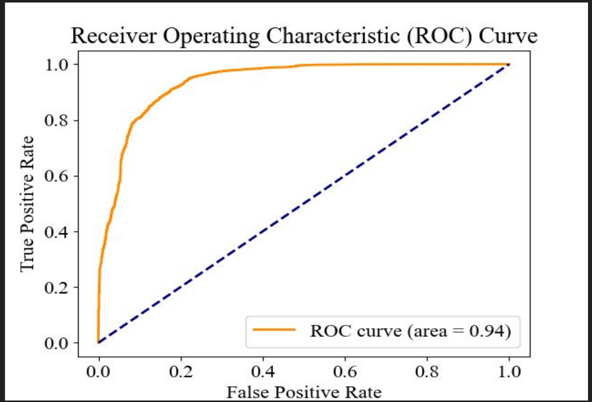
Results

Metrics such as Cohen's Kappa, accuracy, precision and recall for the classification algorithms is tabulated in Table below. The ROC-AUC curve for both is plotted in this Figure

Logistic Regression



Random Forest



Algorithm	**Accuracy**	**Precision**	**Recall**	**Kappa-Score**
Logistic regression	0.885	0.894	0.95	0.71
Random forest	0.894	0.887	0.972	0.73

Plan

- Pre-process the hopper observation data from FAO in different continents for all features and develop a model based on the preprocessed dataset.
- Predict the absence or presence of locust in different cities in South America, Australia, and other continents
- Evaluate the generalizability of machine learning models for different swarming species with different geographical limitations but similar behavioral characteristics like locust swarms.

Limitations

- Adding pseudo-absence points may create bias in predictions
- It's difficult to determine absence of a species during surveys
- We need both presence-only and pseudo-absence models for accurate analysis

References

Diego Gómez et al. “Machine learning approach to locate desert locust breeding areas based on ESA CCI soil moisture”. In: Journal of Applied Remote Sensing 12.3 (2018), pp. 036011–036011.

Diego Gómez et al. “Modelling desert locust presences using 32-year soil moisture data on a large-scale”. In: Ecological Indicators 117 (2020), p. 106655.

Diego Gómez et al. “Prediction of desert locust breeding areas using machine learning methods and SMOS (MIR_SMNRT2) Near Real Time product”. In: Journal of Arid Environments 194 (2021), p. 104599.

Emily Kimathi et al. “Prediction of breeding regions for the desert locust Schistocerca gregaria in East Africa”. In: Scientific Reports 10.1 (2020), p. 11937.

Ibrahim Salihu Yusuf et al. “On pseudo-absence generation and machine learning for locust breeding ground prediction in Africa”. In: arXiv preprint arXiv:2111.03904 (2021)\