Prediction of locust swarms using machine learning

Sunaina 1911182 **Bibhu** 1911052

Abstract

Locust swarming is a behavioural phase transition problem in ecology where the population can be pushed from one alternative stable state to another depending upon the population density. It hovers between swarming and recession. The way this phase transition interacts with its environmental factors is a critical problem to understand as locust swarms decimate crops and pastures in a very short amount of time. This further leads to famines in developing countries and affects the livelihood of local people. In this paper, we implement baseline models to understand locust swarming and the different environmental variables that impact it. We will further extrapolate it to Latin American, India and other countries where there there is a gap of models predicting the locust swarms. Our work will provide insights into the the ecology of locust swarms and generalisability of machine learning models.

1 Introduction

Ecosystems are complex and dynamic systems consisting of both abiotic and biotic parameters that determine its state. Certain ecosystems can undergo rapid change from one state to another. These are called catastrophic regime shifts that has multiple alternative stable states. Empirically, it has been observed that a particular stable state is observed until its "tipping point" is reached and then an alternative stable state is observed. A few examples of rapidly-changing ecological systems are coral bleaching, desertification and locust swarming.

Locusts are found in two stable states - one is solitary where they don't interact with each other and the other is the gregarious state where due to certain environmental conditions, individual locust populations aggregate together to produce huge locust swarms. The tipping point for this system is the critical population density that decides the state of the system. Many mechanistic models based on statistical physics have been worked upon especially related to collective behaviour[1]. But these models are not very helpful to predict the swarming. Machine learning as a tool can be impactful to understand locust swarming from a predictive and theoretical background. This is possible since although the system is dependent upon the tipping point, there are indirect abiotic factors deciding the tipping point. Some of them are the surrounding soil moisture where the locust eggs are present or the temperature of the area where adult locusts are present that can fasten up their metabolism and so on.

Recently, machine learning models have been used to predict locust swarming with high accuracy and precision.[6, 9] Some of the features that have been used are: soil moisture, precipitation, average temperature, soil type and so on. Recent works have been tabulated in Table 1.

2 Related Works

Abiotic parameters as elucidated in the table have been used in the different predictive models. Locust swarming is associated with arid regions and sudden changes in precipitation. As for biotic parameters, hopper presence/absence have been used as the target feature. The life cycle of locusts consists of three stages: egg, nymph (also known as hopper) and adult. Both nymph and adult stages are capable of producing swarms, albeit the nymph swarms are called bands and are smaller than swarms. These bands are also the cause of crop loss and habitat destruction. The tipping point of

Table 1: Related works

Papers	Countries used	Features used
Prediction of breeding regions for the Desert Locust *Schistocerca Gregaria* in East Africa.[5]	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. [4]	30 countries	Soil moisture for prediction of nymph population
Modelling Desert Locust presences using 32-year soil moisture data on a large-scale.[3]	30 countries	Soil moisture for prediction of nymph population.
Machine learning approach to locate desert locust breeding areas based on ESA CCI soil moisture [2]	Mauritania	Soil moisture for prediction of nymph populations
On pseudo-absence generation and machine learning for locust breeding ground prediction in Africa [10]	African countries	Soil moisture (at different depths), average temperature, wind, rainfall and quality of air.*

locust swarms is dependent upon the number of nymph locusts that are turned into swarming adults. Most studies have used this as the target label from a preventative aspect although the presence of eggs and adult locusts can also be used.[7]

2.1 Methodology followed

2.1.1 Pre-processing and Feature Engineering

Time series data from 1985-2021 is collected from Food and Agriculture Organization's locust swarming dataset comprising the hopper absence or presence at different coordinates all over the world through its global Desert Locust Information Service. Considering that we are trying to predict the hopper population, the time series data from 95 days prior from the time when the presence data was collected is scraped for different environmental variables and different statistical descriptions (mean, median, maximum, minimum) can be used to engineer new features. 95 days is the maximum amount of number at which eggs are laid and develop into larvae to produce hoppers. From -95 to 0, further buckets are made of different time intervals such as 6, 12, 16 and so on. Over this interval, the different statistical features are calculated. For example, in a 6-day bucket, you may have a feature such as Tavg _95-89 which is an average of the temperatures between Day 95 and Day 89. It has been observed that the smaller this interval is, the higher the accuracy of the model as detailed by Gómez et al. [3]

For all these intervals, for each X and Y coordinate the temperature, precipitation, soil moisture and other environmental variables are scraped from various meteorological satellite datasets such as GLDAS Noah Land Surface Model (0.25 x 0.25), LANDSAT, etc.

These features undergo suitable pre-processing steps such as centering and scaling as the ranges for different features are different. The model is trained on one set of countries and tested on another set of countries.

2.2 Pseudo-generation of absence points

It’s difficult to ascertain the absence of a species in an area during ecological surveys. To deal with this, researchers generate pseudo-absence points near the presence zones. There are a number of ways to perform this as reviewed in a paper (Yusuf et al., 2022). Two popular methods are to either randomly provide absence points or through environmental profiling where environmental variables of the nearby regions are also taken into consideration. The absence points are important for feeding datasets in machine learning models so that there is not an overrepresentation of one class in the data. Although most machine learning models perform better on datasets with absence points with low bias, there are also machine-learning based species distribution models (SDM) that use presence-only data. One such popular SDM model is MaxEnt. [8] There’s a caveat though - MaxEnt still generates "background" points but it doesn’t associate these points with the absence of the species. MaxEnt aims to map the optimal environmental parameters with the presence of the species.

2.3 Models used and their results

Different machine learning models such as logistic regression, k-Nearest Neighbors, MaxEnt, XGBoost have been used in the literature. Depending on the countries and features used to test their model on, they get varied results. Environmental variables such as soil moisture is a good predictor even when it is used without any other variable. Their results have been tabulated in Table- 3.

Table 3: Comparison of different models

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

3 Baseline algorithms

The baseline algorithms to be implemented are logistic regression and random forest.

4 Experiments

4.1 Curation of dataset

For the dataset, we have used the pre-processing pipeline available from previous works to get the data of African countries from Food and Agricultural Organisation (FAO)’s hopper observation data. All the inexact entries were removed. The X and Y coordinates from this data were used to fetch the data from GLDAS Noah Land Surface Model and SoilGrids for 95 days prior the presence data was collected as performed by Yusuf et al. [10]. Further bucketizing based on time interval of 6 days created total 1168 features. The dataset information is shown in Table- 5.

For training and testing, the entire dataset of all the different countries and timeline was split into two subsets with test size chosen to be 0.34.

Table 5: Dataset

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

4.2 Parameters

Two baselines were chosen to be implemented – regularised logistic regression with the default L2 penalty term and random forest. Other than the features that have been described above, the hyperparameters of the models were tuned. For logistic regression, the number of iterations were tuned and for random forest, the number of trees were tuned. The related codes can be found in the following github repository.

4.3 Results

Metrics such as Cohen's Kappa, accuracy, precision and recall for the classification algorithms is tabulated in Table- 7. The ROC curve for both is plotted in Figure- 1.

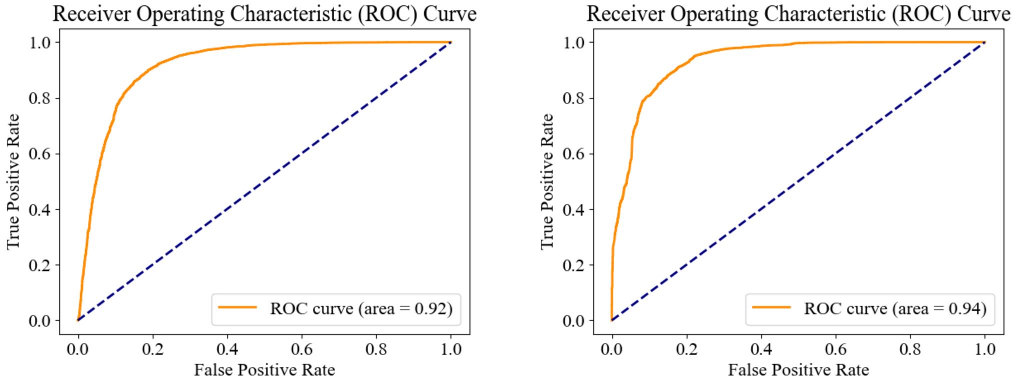

Figure 1: Figure 1: ROC curve of logistic regression and random forest respectively

Table 7: Classification metrics

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

5 Plan

Now that we have a baseline model for the African countries, we plan to pre-process the entire data of FAO's hopper observation data in different continents. This has been performed for the soil moisture feature to some extent but we plan to perform it for all the others. Once we have a model based on this dataset, we will predict the absence or presence of locust in different cities in South America, Australia and so on. This will help us understand the generalizability of machine learning models for different swarming species with different geographical limitations but similar behavioural characteristics such as the formation of locust swarms.

6 Limitations

Addition of pseudo-absence points may be creating a bias while prediction – considering that the test data has absence data points that has been generated so the high accuracy and precision of the baseline models is also due to the correct prediction of those points. Ecologically, it is very hard to conclude that. But since most of the classification algorithms require a similar distribution for both classes in the dataset that it is training on, we still need to generate these points. For this reason, we need both pseudo-absence and presence-only models for any predictive analysis. Thus, we need to implement MaxEnt species distribution model as well.

References

[1] Leah Edelstein-Keshet, James Watmough, and Daniel Grunbaum. “Do travelling band solutions describe cohesive swarms? An investigation for migratory locusts”. In: *Journal of mathematical biology* 36 (1998), pp. 515–549.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Karthika Suresh Kumar and Aamer Abdul Rahman. “Early detection of locust swarms using deep learning”. In: *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019*. Springer, 2020, pp. 303–310.

[7] Marion Le Gall, Rick Overson, and Arianne Cease. “A global review on locusts (Orthoptera: Acrididae) and their interactions with livestock grazing practices”. In: *Frontiers in Ecology and Evolution* 7 (2019), p. 263.

[8] Steven J Phillips et al. “Opening the black box: An open-source release of Maxent”. In: *Ecography* 40.7 (2017), pp. 887–893.

[9] Ruiqi Sun et al. “Dynamic Forecast of Desert Locust Presence Using Machine Learning with a Multivariate Time Lag Sliding Window Technique”. In: *Remote Sensing* 14.3 (2022), p. 747.

[10] 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).