# Evaluation of Machine Learning Models for Species Distribution Modeling in the Amazon

Renato Miyaji[1], Felipe De Almeida[1], and Pedro Corrêa[1]

[1]University of São Paulo

November 24, 2022

**Abstract**

Species Distribution Modelling (SDM) is widely used by ecologists to monitor biodiversity and manage wildlife. In the last decades, Artificial Intelligence (AI) and Machine Learning (ML) techniques became popular and were successfully applied for different tasks, including SDM. The objective of this article was to evaluate Machine Learning models for Species Distribution Modeling in the Amazon Basin region near Manaus (AM), based on meteorological and aerosol data collected by the GoAmazon 2014/15 project. The techniques were evaluated regarding their accuracy and the Decision Tree Classifier and the Maximum Entropy Model obtained good predictive performances.

# Evaluation of Machine Learning Models for Species Distribution Modeling in the Amazon

MIYAJI, R. O[1] ; DE ALMEIDA, F. V.[1]; CORRÊA, P. L. .P.[1]

[1] Escola Politécnica - Universidade de São Paulo

**Email**: {re.miyaji, felipe.valencia.almeida, pedro.correa }@usp.br

## INTRODUCTION

Species Distribution Modelling (SDM) is widely used by ecologists to monitor biodiversity and manage wildlife. In the last decades, Artificial Intelligence (AI) and Machine Learning (ML) techniques became popular and were successfully applied for different tasks, including SDM. Since the beginning of the century, hundreds of articles were published regarding the use of AI or ML techniques for SDM.

## OBJECTIVES

The objective of this article was to evaluate Machine Learning models for Species Distribution Modeling in the Amazon Basin region near Manaus (AM), based on meteorological and aerosol data collected by the GoAmazon 2014/15 project. The techniques were evaluated regarding their accuracy.

## MATERIALS AND METHODS

The methodology used for the Species Distribution Modeling was adapted from [1]. It is composed of five steps: construction of the scientific hypothesis, data pre-analysis, modeling, prediction, and validation of the scientific hypothesis.

For this article, the scientific hypothesis was defined as that the variation in the concentration of aerosols, due to anthropic action, influences the distribution of species in the region of Manaus and Manacapuru (AM).

A study case in the Amazon Basin was selected to test this hypothesis, because the city of Manaus is considered by specialists as an ideal laboratory to study the effects of the anthropic action in a tropical Rainforest ecosystem [2]. A bioclimatic dataset was necessary to develop the Species Distribution Modeling experiment. Meteorological and aerosol data generated by spatial interpolations provided by [3] were collected. In this dataset, there were 10 different variables: temperature, concentrations of ozone (O3), carbon monoxide (CO), nitrogen oxide (NOX), methane (CH4), carbon dioxide (CO2), isoprene and acetonitrile, numerical concentration of particles and water volumetric fraction (H2O). All variables were originally collected in low altitude flights performed by an aircraft during the GoAmazon 2014/15 project.

In addition, species occurrence data was also collected in the repository of the Biodiversity Portal of the Chico Mendes Institute for Biodiversity Conservation (ICMBio 2021). The data treatment aimed to obtain a bioclimatic dataset that allowed the development of Species Distribution Modeling. From this dataset, it was necessary to define the species to be analyzed: the black vulture (Coragyps atratus).

In the modeling step, the predictor variables were selected, through Correlation Analysis, to avoid the occurrence of multicollinearity.

Then, the Species Distribution Modeling algorithm was selected. Two techniques were compared: the Maximum Entropy Model (MaxEnt), which is one of the most common for Species Distribution Modeling and has a good potential for datasets of small and medium dimension [4], and a Machine Learning technique, the Decision Tree Classifier (DT).

After applying these techniques, their performance was evaluated and compared, considering AUC-ROC (Area Under the Receiver Operating characteristic Curve) and classification metrics, such as accuracy, precision, and recall. All metrics were calculated over the Hold-out dataset.

Finally, the scientific hypothesis could be evaluated with the results obtained by the fitted models.

| Model | Accuracy | Recall | Precision | ROC-AUC |
|---|---|---|---|---|
| DT | 99 % | 50 % | 1 % | 83 % |
| MaxEnt | 75 % | 85 % | 0 % | 80 % |

TABLE I
CLASSIFICATION METRICS IN THE HOLD-OUT DATASET

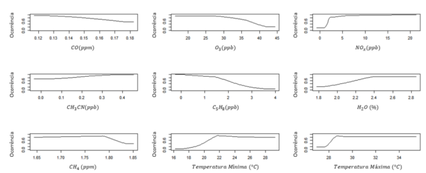

Fig. 1. Response Curves for *Coragyps atratus*

## RESULTS

The performance metrics of the models are presented in Table I.

Both models obtained a high accuracy, with almost 99 % for the Decision Tree Classifier. Considering recall, the best model was the Maximum Entropy Model with 85%. The ROC-AUC was high and almost the same for both models: 83% for the Decision Tree Classifier and 80% for Maximum Entropy Model.

The results indicate that the models had a good predictive performance and were able to increase the True Positives (VP) and reduce the False Negatives (FN), leading to a high recall, but the False Positives (FP) remained, leading to a very low precision.

To validate the scientific hypothesis, the response curves from the Maximum Entropy Model, presented in Figure 1, could be analyzed. For the Maximum Entropy Model, the variables that had the biggest negative effect on the probability of occurrence of Coragyps atratus were the concentration of isoprene, ozone and methane.

For the Decision Tree Classifier, the scientific hypothesis could be evaluated based on the decision tree that was fitted. The variables that were used for the first conditions of the decision tree were: the concentration of isoprene, nitrogen oxide and carbon monoxide. These are considered the most relevant by the model, considering information gain.

It was also possible to obtain partial dependence plots for the Decision Tree Classifier, which are similar to the response curves of the Maximum Entropy Model. For the Decision Tree Classifier, the concentration of isoprene and ozone had the biggest negative effect on the occurrence of Coragyps atratus.

## CONCLUSIONS

In conclusion, it was possible to develop a Species Distribution Modeling experiment with a case study in the Amazon Basin and evaluate Machine Learning models. Two different models were compared: the Maximum Entropy Model and the Decision Tree Classifier. The best model for ROC-AUC was the Decision Tree with 83%, considering recall, the best was the Maximum Entropy with 85%. With both models, it was possible to evaluate the scientific hypothesis that the variation in the concentration of aerosols, due to anthropic action, influences the probablity of occurrence of Coragyps atratus in the region of Manaus and Manacapuru (AM).

**References**:

[1] PINAYA, J.; CORRÊA, P. Metodologia para definição das atividades do processo de modelagem de distribuição de espécies. In: Anais do V WCAMA. Porto Alegre, RS, Brasil, 2014. p. 45–54.

[2] MARTIN, S. T. et al. The green ocean amazon experiment (goamazon2014/5) observes pollution affecting gases, aerosols, clouds, and rainfall over the rain forest. Bulletin of the American Meteorological Society, v. 98, n. 5, p. 981–997, 2017.

[3] MIYAJI, R. O. et al. Spatial interpolation of air pollutant and meteorological variables in central amazonia. Data, v. 6, n. 12, 2021.

[4] PHILLIPS, S. J. Maximum entropy modeling of species geographic distribution. Ecological Modelling, v. 190, p. 231–259, 2005.

**Acknowledgements:**

The authors would like to thank CNPq for the financial support and to ARM and FAPESP for the data availability under projects (2017/ 17047-0 and 2020/ 15230-5).

**Author affiliations:** Renato Okabayashi Miyaji (University of São Paulo, BR); Felipe Valencia de Almeida (University of São Paulo, BR); Pedro Luiz Pizzigatti Corrêa (University of São Paulo, BR).