# Experimento de modelagem de distribuição de espécies baseada em variáveis ambientais e de aerossóis na região próxima a Manaus (AM)

**Felipe V. de Almeida**[1], **Weslley M. Bueno**[1], **Renato O. Miyaji**[1], **Pedro L. P. Corrêa**[1]

[1]Escola Politécnica – Universidade de São Paulo (USP)

{felipe.valencia.almeida,weslley.bueno,re.miyaji,pedro.correa}@usp.br

***Abstract.*** *The Amazon Rainforest region is considered to be the one that concentrates the greatest biodiversity in the world. Seen by many as a unique study laboratory, there are numerous possibilities for applications of computational models to extract value from data collected in the region. This work presents an experiment of modeling the distribution of species located in a region close to Manaus (AM). This modeling was based on environmental and aerosol variables, whose raw data were obtained from the GOAmazon project repository. The integration of environmental data with species occurrence data of Brazilian fauna allowed the models described to predict the probability of a species being present in the studied area.*

***Resumo.*** *A região da Floresta Amazônica é considerada como sendo a que concentra a maior biodiversidade do mundo. Visto por muitos como um laboratório único de estudos, são inúmeras as possibilidades de aplicações de modelos computacionais para extrair valor de dados colhidos na região. Este trabalho apresenta um experimento de modelagem de distribuição de espécies localizadas em região próxima a Manaus (AM). Esta modelagem foi baseada em variáveis ambientais e de aerossóis, cujos dados brutos foram obtidos em repositório do projeto GOAmazon. A integração de dados ambientais com dados de ocorrência de espécies da fauna brasileira permitiu aos modelos descritos prever a probabilidade de uma espécie estar presente na área estudada.*

## 1. Introdução

No cenário atual a Floresta Amazônica está em constante debate pela comunidade mundial. Seja pela sua extensa cobertura vegetal ou pelo fato de concentrar parcela significativa da biodiversidade do globo, sua importância para a comunidade científica é inquestionável. Atrelado a isso, dentre os 17 Objetivos de Desenvolvimento Sustentável (ODS) lançados pela Organização das Nações Unidas (ONU) para a agenda 2030 dois estão relacionados com esta. São eles: ”13 - Ação contra a mudança global do clima” e “15 - Vida terrestre” tendo este último uma aderência maior com o escopo deste trabalho.

No período de 2014 a 2015 foi realizado um projeto denominado *GOAmazon*, através de uma parceria entre o *Atmospheric Radiation Measurement* (ARM), laboratório vinculado ao Departamento de Energia dos Estados Unidos, e instituições brasileiras como a USP, a UEA e o INPE. O projeto em questão teve como intuito identificar a relação entre vegetação, variáveis atmosféricas e a produção de aerossóis [Martin et al. 2016]. Para tal, foram realizadas diversas medições de parâmetros climáticos distintos no período

de 2 anos do projeto. Os dados foram obtidos a partir de medições de nove estações de pesquisa localizadas em regiões próximas ao município de Manaus (AM) e de voos realizados por aviões equipados com sensores de medição também em região próxima. Os voos em questão foram realizados em dois períodos de operação intensa (IOPs), sendo estes durante a estação seca e a chuvosa.

O seguinte trabalho apresenta um experimento de modelagem de distribuição de espécies da fauna nacional baseada em variáveis colhidas do *dataset* em questão. Os modelos de distribuição de espécies são de grande importância para a ecologia, visto que permitem o monitoramento e a análise da biodiversidade, possibilitando estimar os impactos de mudanças climáticas nas espécies [Elith and Leathwick 2009]. Através de sua aplicação, pode-se identificar as variáveis ambientais que possuem maior influência na ocorrência das espécies, além de determinar seu nicho ecológico [Hutchinson 1991].

Inicialmente são apresentados os trabalhos relacionados ao tema, seguido pela metodologia aplicada no contexto do trabalho e os resultados obtidos. Ao término são apresentadas as considerações finais e oportunidades de trabalhos futuros.

## 2. Trabalhos Relacionados

Visando localizar na literatura trabalhos que relacionem os dados do *GOAmazon* com a modelagem de distribuição de espécies, foi realizada uma pesquisa nas grandes bases indexadoras de artigos científicos (Elsevier, ACM, IEEE e Springer). Optou-se por uma *query* simples de pesquisa, sendo esta apenas "*GOAmazon*". Devido ao baixo número de artigos encontrados (32 artigos), não foi necessário realizar um tratamento mais sofisticado na *query*.

Observou-se que uma parcela majoritária dos artigos obtidos nesta busca situam-se em congressos/periódicos voltados para a área das Ciências Atmosféricas. Alguns trabalhos são voltados para uma análise geral dos dados climáticos de aerossóis obtidos, como o trabalho de [Cirino et al. 2018], enquanto outros possuem enfoque específico para um conjunto reduzido de variáveis coletadas, como o trabalho de [Wei et al. 2019].

Se tratando exclusivamente do problema da modelagem de distribuição de espécies, encontra-se na literatura alguns trabalhos relacionados. Dentre eles destaca-se aqui o trabalho de [Pinaya and Corrêa 2014], que apresenta uma metodologia para realização de um experimento de modelagem de distribuição de espécies e o de [Carneiro et al. 2016] que apresenta uma análise das limitações do uso da modelagem de distribuição de espécies em um estudo de caso na Amazônia. Também são encontrados outros trabalhos com o propósito de realizar experimentos semelhantes ao aqui proposto, porém com o uso de *datasets* distintos e em regiões diferentes. [Effrosynidis et al. 2020] por exemplo afirma apresentar a primeira aplicação do algoritmo *Extremely Gradient Boosting* (*XGBoost*) no contexto do problema da modelagem de distribuição de espécies, onde o desempenho deste algoritmo é visto como o melhor dentre os demais analisados.

Desta forma, não foi encontrado em nenhum dos artigos em questão a tentativa de relacionar os dados do projeto *GOAmazon* com a modelagem de distribuição de espécies.

## 3. Metodologia

A metodologia para a realização do experimento é apresentada em etapas pela Figura 1. A seguir são apresentadas considerações relacionadas às etapas destacadas.

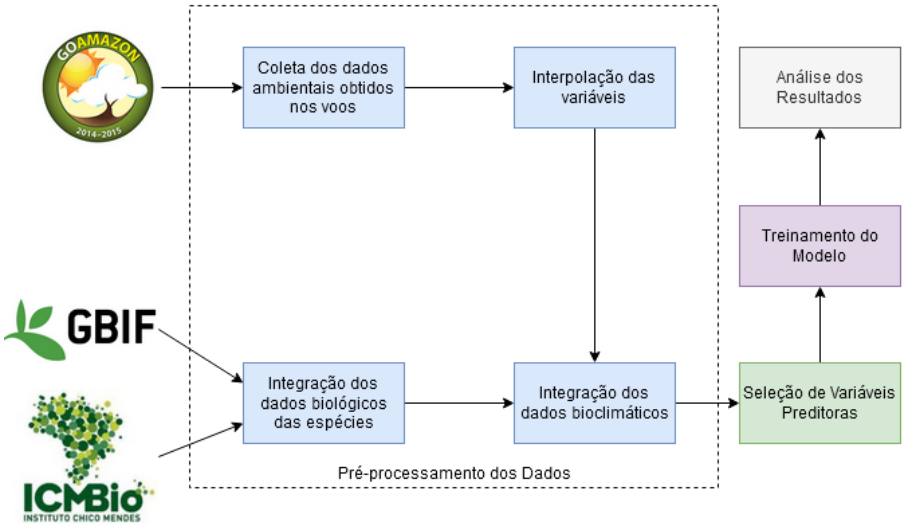



**Figura 1. Metodologia adotada no trabalho**

### 3.1. Pré-processamento dos Dados

Os dados utilizados pelo trabalho provém de três fontes principais. Os dados ambientais e de aerossóis foram obtidos a partir dos repositórios do projeto *GOAmazon*. Esses foram coletados por meio de uma aeronave, que realizou 35 voos distribuídos em dois períodos de operação, sendo o escopo deste trabalho os dados correspondentes ao que ocorreu durante a estação seca. As variáveis consideradas foram: a temperatura, as concentrações de monóxido de carbono ($CO$), ozônio ($O_3$), óxidos de nitrogênio ($NO_X$), dióxido de carbono ($CO_2$), metano ($CH_4$), compostos orgânicos voláteis, como o isopreno e a acetonitrila, a concentração numérica de partículas (CPC) e a fração volumétrica de água ($H_2O$).

De modo a expandir a área disponível das variáveis ambientais e de aerossóis, aplicou-se uma metodologia de interpolação espacial, utilizando a técnica de interpolação linear baricêntrica segmentada, sendo a com menor erro quando comparada com as demais. Essa foi aplicada para cada variável coletada em cada um dos voos realizados na estação seca. A partir das superfícies interpoladas para cada voo, foi feita uma sumarização para cada variável. Dessa forma, obteve-se superfícies médias de interpolação. O tratamento e a manipulação dos dados foi realizada na linguagem Python, através da aplicação *web Jupyter Notebook* (Jupyter Team, 2015).

Já os dados de ocorrência de espécies foram coletados a partir de dois repositórios: do Portal de Dados da Biodiversidade do Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio) e do *Global Biodiversity Information Facility* (GBIF). Aplicou-se um filtro para os dados, de modo que eles fossem referentes à mesma região e à mesma data dos dados ambientais e de aerossóis, sendo necessário ajustar a precisão das coordenadas de latitude e longitude, de acordo com o passo utilizado na construção da malha de interpolação. No total, foi possível obter registros de ocorrência de 40 espécies distintas na estação seca, sendo 39 delas pertencentes à classe Aves e uma delas à classe *Reptilia*. Para que uma espécie fosse considerada, adotou-se o critério que ela deveria possuir, no mínimo, 17 ocorrências, conforme recomendado por [Pinaya and Corrêa 2014]. A Tabela 1 sumariza as ocorrência das espécies na área em questão.

| espécie | nº de ocorrências | espécie | nº de ocorrências | espécie | nº de ocorrências |
|---|---|---|---|---|---|
| *Ammodramus aurifrons* | 31 | *Aratinga leucophthalma* | 27 | *Ardea alba* | 18 |
| *Brotogeris sanctithomae* | 32 | *Brotogeris sanctithomae* | 24 | *Brotogeris sanctithomae* | 17 |
| *Brotogeris sanctithomae* | 22 | *Cacicus cela* | 34 | *Caiman crocodilus* | 01 |
| *Cathartes aura* | 23 | *Columbina_passerina* | 19 | *Columbina talpacoti* | 19 |
| *Coragyps atratus* | 54 | *Crotophaga ani* | 33 | *Crotophaga major* | 18 |
| *Dendrocygna autumnalis* | 21 | *Jacana jacana* | 27 | *Megaceryle torquata* | 23 |
| *Milvago chimachima* | 38 | *Myiozetetes cayanensis* | 18 | *Pandion haliaetus* | 18 |
| *Patagioenas cayennensis* | 17 | *Phaetusa simplex* | 30 | *Phalacrocorax brasilianus* | 17 |
| *Pitangus sulphuratus* | 41 | *Ramphocelus carbo* | 24 | *Rostrhamus sociabilis* | 19 |
| *Rupornis magnirostris* | 18 | *Sicalis columbiana* | 25 | *Sporophila castaneiventris* | 18 |
| *Stelgidopteryx ruficollis* | 20 | *Sturnella militaris* | 17 | *Tachycineta albiventer* | 22 |
| *Thraupis episcopus* | 41 | *Thraupis palmarum* | 31 | *Todirostrum maculatum* | 31 |
| *Troglodytes aedon* | 31 | *Turdus leucomelas* | 18 | *Tyrannus melancholicus* | 50 |
| Tyrannus savana | 18 | | | | |

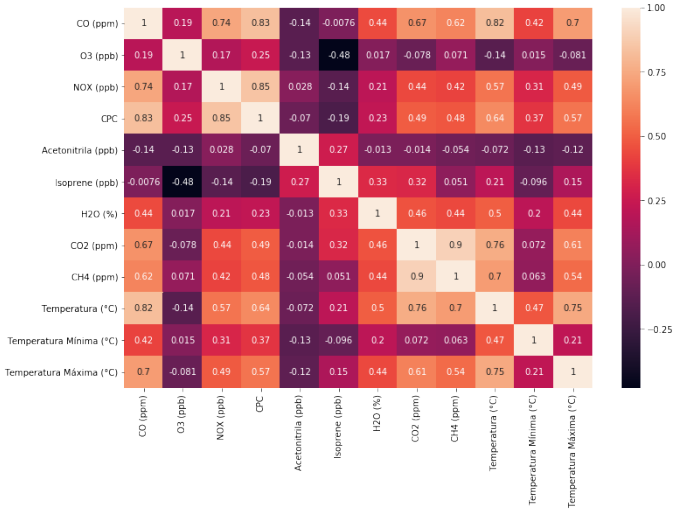
**Tabela 1. Ocorrência das espécies na área interpolada**

Desse modo, construiu-se um *dataset* com dados bioclimáticos para a estação seca, sobre o qual seria possível aplicar a modelagem de distribuição de espécies. A construção desse foi feita por meio da operação de junção do *dataset* de dados ambientais e de aerossóis com o de dados de ocorrência de espécies, considerando as colunas de georreferenciamento (latitude e longitude) e de data de registro da ocorrência, de modo a se obter um *dataset* final, que possuía as coordenadas de latitude e longitude como chave primária composta.

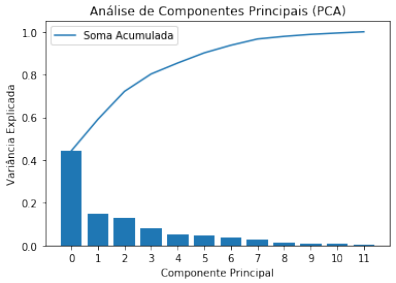
### 3.2. Seleção de Variáveis Preditoras

Para selecionar as variáveis a serem utilizadas como preditoras para o modelo de distribuição de espécies, analisou-se a correlação existente entre elas. Para tal adotou-se o coeficiente de Pearson, através do qual foi possível quantificar a relação linear existente entre as variáveis. Essa análise teve como objetivo evitar que o modelo desenvolvido perdesse qualidade, através da incorporação de padrões aleatórios entre as variáveis, e que o fenômeno de multicolinearidade ocorresse [Pinaya and Corrêa 2014]. Analisou-se a matriz de correlação para o *dataset* apresentada na Figura 2a, retirando uma das variáveis dos pares que eram muito correlacionados entre si – com coeficiente de Pearson com módulo superior a 80% [Mateo et al. 2013]. Assim, optou-se por retirar as variáveis concentração numérica de partículas (CPC), concentração de dióxido de carbono e temperatura.

Ademais, também realizou-se uma análise de componentes principais (PCA), com o objetivo de se reduzir a dimensão do *dataset*. Através dessa, foram calculados os autovalores da matriz de correlação, que representam os componentes principais, cujos carregamentos apresentam uma medida da importância de cada variável para variância total. Por meio dessa, buscou-se que o modelo não desenvolvesse padrões excessivamente complexos, de modo a não tender ao fenômeno de *overfitting* [Pinaya 2019]. A variância explicada para cada componente principal foi avaliada. A partir da Figura 2b, notou-se, pela soma acumulada da variância, que os nove primeiros componentes principais representavam aproximadamente 95% da variância explicada total do *dataset*. Assim, praticamente não haveria perda ao se reduzir a dimensão do *dataset*. Portanto, foi possível confirmar a alteração proposta através da análise da matriz de correlação: retirando as variáveis altamente correlacionadas, obteve-se um *dataset* a ser utilizado com dimensão igual a nove.




**(a) Matriz de correlação para variáveis do *dataset*.**




**(b) Variância explicada para variáveis do *dataset* através da análise de componentes principais (PCA).**

**Figura 2. Matriz de correlação e análise de componentes principais.**

### 3.3. Treinamento do Modelo

Antes do treinamento dos modelos, eliminou-se do *dataset* as linhas que continham ao menos um valor "NaN"(*Not a Number*) para as variáveis ambientais estudadas, reduzindo o tamanho do *dataset* de 1.032.031 amostras para 183.331 amostras, uma redução de aproximadamente 83%. Essa etapa foi feita de modo a garantir que todas as variáveis preditoras possuíssem valor diferente de "NaN"em cada linha do *dataset* a ser utilizado, condição necessária para a aplicação de modelos de classificação, como o de regressão logística [Scikit-learn Developers 2020].

Ademais, estabeleceram-se os casos de teste. Estes foram definidos em função das dimensões da região geográfica sob análise. Utilizou-se em um caso toda a região não-nula do *dataset* e em outro restringiu-se a região ao menor quadrilátero que continha todas as amostras positivas para uma dada espécie. A escolha por essa variação justifica-se pelo fato de que, geralmente, modelos de distribuição de espécies tendem a apresentar uma acurácia maior quando aplicados a um *dataset* com uma proporção de pontos de ocorrência de espécies maior em relação ao seu tamanho total [Hernandez et al. 2006].

Das 40 espécies presentes no *dataset*, foi selecionada a espécie *Coragyps atratus* para apresentação neste trabalho por conter o maior número de amostras positivas em comparação com as outras. Além disso, as ocorrências da espécie não eram espacialmente esparsas, fato que contribui para o aumento da acurácia do modelo [Hernandez et al. 2006]. Considerando a região completa, a porcentagem de amostras positivas da *Coragyps atratus* antes do balanceamento é de 0,029%, enquanto que na região restrita a porcentagem é de 0,046%. Apesar de ser uma proporção baixa, é a maior encontrada entre todas as 40 espécies para os dois cenários.

Dado que as proporções entre amostras positivas e negativas das espécies atingiam a ordem de $10^{-4}$, a tarefa configurou-se como sendo de *classificação desbalanceada*. Para a modelagem de distribuição de espécies, essa condição dificulta o desenvolvimento de classificadores que apresentem acurácia elevada [Johnson et al. 2012]. Assim, foram aplicadas técnicas de *resampling* para aumentar a proporção de classes po-

sitivas. Para isto, dividiu-se o conjunto inicial em um conjunto de testes e de treino na proporção 30/70, de forma estratificada, aplicando-se então as técnicas de balanceamento sobre o conjunto de treino. Para aumentar o número de amostras positivas, utilizou-se a técnica *Synthetic Minority Oversampling Technique* (SMOTE) para criação de amostras sintéticas positivas [The Imbalanced-learn Developers 2021]. A aplicação dessa técnica mostrou-se ser uma das mais efetivas para tratar o contexto de classificação desbalanceada [Johnson et al. 2012]. Os parâmetros utilizados para o balanceamento foram definidos para que o *dataset* resultante contivesse uma proporção de 1:3 entre amostras positivas e negativas.

Foram aplicados e comparados dois modelos de aprendizagem de máquina de classificação para a tarefa da modelagem de distribuição de espécies. O primeiro deles, um modelo baseado na técnica de *Gradient Boosting* [Friedman 2002] construído a partir de classificadores base (árvores de decisão), o *Extremely Gradient Boosting* (*XGBoost*) [XGBoost Developers 2020]. Trata-se de um modelo emergente e praticamente inédito no contexto de modelagem de distribuição de espécies, mas que apresentou desempenho superior em comparação com outros no trabalho desenvolvido por [Effrosynidis et al. 2020]. O segundo modelo utilizado foi a regressão logística, um dos mais recorrentes para a modelagem de distribuição de espécies [Berg et al. 2004].

O experimento com ambos os modelos foi feito na linguagem Python, utilizando as funções existentes nas bibliotecas dedicadas à aplicações de aprendizagem de máquina, *Scikit-learn* [Scikit-learn Developers 2020] para a regressão logística e *XGBoost* [XGBoost Developers 2020].

A configuração dos hiperparâmetros do *XGBoost* foi feita utilizando-se a função `GridSearchCV()` da biblioteca *Scikit-learn*, que compara o desempenho de diferentes modelos definidos a partir da combinação de valores dos hiperparâmetros definidos em grades, por meio de uma validação cruzada [Scikit-learn Developers 2020]. Para a validação cruzada, utilizou-se o algoritmo de *3-folds*. Verificou-se que o parâmetro `n_estimators`, que define o número de árvores de decisão empregadas, foi o mais sensível em causar *overfitting* no modelo. A princípio realizou-se uma busca em um intervalo de valores altos, entre 300 e 1000, mas posteriormente verificou-se que os valores ideais se encontravam em intervalos baixos, entre 10 e 20 árvores, reduzindo a complexidade do modelo. A taxa de aprendizado foi outro parâmetro que necessitou de ajuste fino para otimização do modelo. Os valores utilizados nos parâmetros foram: `n_estimators = 3`, `max_depth = 4`, `min_child_weight = 2` e `learning_rate = 0.05`.

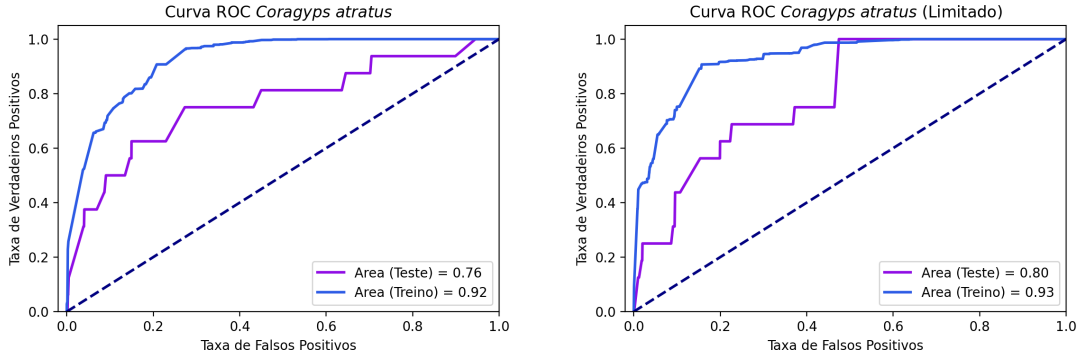
Na regressão logística o parâmetro *C*, que corresponde ao inverso da força de regularização [Scikit-learn Developers 2020], controla a robustez do modelo a pequenas variações dos dados, tornando-o mais resistente ao fenômeno de *overfitting*. Assim, adotou-se um valor pequeno para este parâmetro (`C=0.01`). Para os demais parâmetros, utilizou-se os valores padrões fornecidos pela biblioteca.

Foram treinados os dois modelos, correspondendo à região completa e não-nula do *dataset* e à região restrita. Os modelos foram avaliados através da área embaixo da curva (AUC) característica de operação do receptor (ROC), utilizando-se os valores previstos pelos modelos quando aplicados sobre o conjunto de testes. Além disso, comparou-se os

valores com os obtidos no conjunto de treinamento para avaliar a ocorrência do fenômeno de *overfitting*.

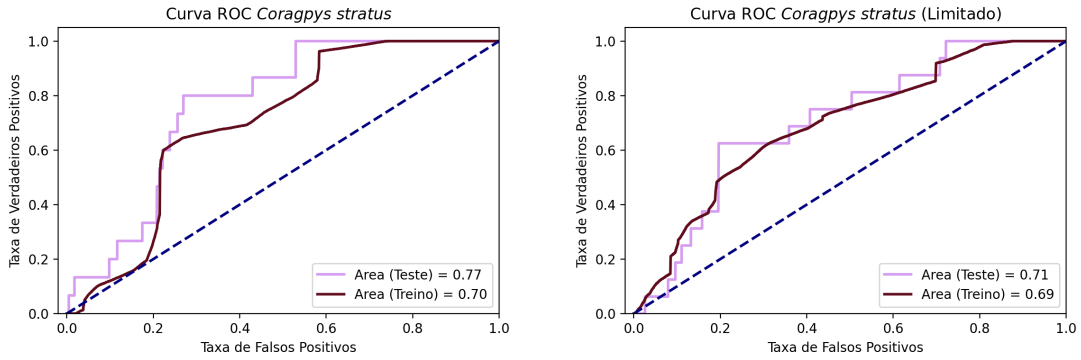
## 4. Resultados e Discussões

Em números absolutos, os resultados obtidos pelos dois modelos, em ambos os cenários (região completa e restrita), são próximos. As Figuras 3 e 4 apresentam, respectivamente, as curvas ROC dos modelos gerados utilizando-se o *XGBoost* e a regressão logística para as regiões completa e restrita. Para o *XGBoost* nota-se que o melhor resultado foi obtido utilizando-se a região restrita, ilustrado na Figura 3b, com ROC AUC de 0.8. Para a regressão logística, o melhor resultado foi obtido utilizando-se a região completa, como ilustrado na Figura 4a, com ROC AUC de 0.77.




**(a) Região completa não-nula** **(b) Região restrita não-nula**

**Figura 3. Curvas ROC dos resultados obtidos pelo algoritmo XGBoost para a espécie *Coragyps atratus*.**




**(a) Região completa não-nula** **(b) Região restrita não-nula**

**Figura 4. Curvas ROC dos resultados obtidos pelo algoritmo de regressão logística para a espécie *Coragyps atratus*.**

Utilizou-se então os modelos das Figuras 3b e 4a para prever a probabilidade de ocorrência da espécie *Coragyps atratus* em todo o *dataset*. Os resultados obtidos são apresentados na Figura 5, que apresenta o mapa de distribuição potencial da espécie por toda a região contida no *dataset* e que inclui as cidades de Manaus e Manacapuru (AM).

No entanto, as figuras também apresentam as curvas ROC dos modelos quando aplicados sobre o *dataset* de treinamento, ilustradas pelas curvas roxas na Figura 3 e

pelas curvas magenta na Figura 4. Nota-se que as curvas de treino e teste dos modelos gerados por regressão logística são similares, enquanto que para os modelos gerados pelo *XGBoost* há uma diferença maior entre as duas, o que indica a ocorrência do fenômeno de *overfitting* [Radosavljevic and Anderson 2014].

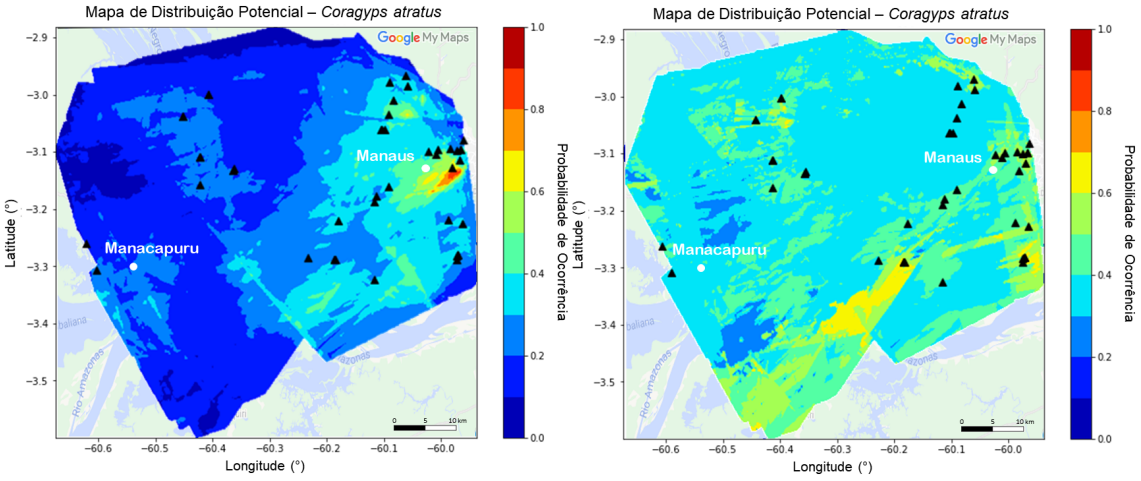



**Figura 5. Mapa de Distribuição Potencial para a espécie *Coragyps atratus* obtido pelo modelo de regressão logística (esquerda) e *XGBoost* (direita). Triângulos em preto representam os pontos de presença observados.**

O modelo apresentado neste trabalho compartilha de algumas das limitações comumente encontradas em outros modelos de distribuição de espécies [Leidenberger et al. 2015], como desbalanceamento entre amostras de presença-ausência da espécie, viés temporal e espacial. Além disso, há o problema de garantir que uma dada espécie não esteja de fato presente em uma dada região, ao contrário de apenas não ter sido observada. Este problema é abordado em [Hegel et al. 2010] onde é realizada uma discussão entre a ocorrência absoluta e a pseudo-ocorrência de uma espécie em determinada área. Enquanto os dados disponíveis nos *datasets* apontam ocorrências absolutas das espécies, identificadas por pesquisadores, os resultados dos modelos treinados apresentam uma malha de probabilidades, que podem ou não indicar a presença de uma espécie em determinada região.

Além disso, mesmo com as técnicas de balanceamento, a acurácia obtida para a presença das espécies ainda foi baixa quando comparada com as amostras negativas, que são majoritárias no *dataset*. Também há as limitações dos próprios algoritmos utilizados. O *XGBoost* é um algoritmo complexo, com muitos hiperparâmetros, o que torna difícil encontrar o conjunto de parâmetros ótimos para um dado cenário. Ressalta-se que cada espécie irá possuir um conjunto distinto, capaz de maximizar a acurácia do modelo para sua ocorrência. Ressalta-se que cada espécie irá possuir um conjunto distinto, capaz de maximizar a acurácia do modelo para sua ocorrência.

## 5. Conclusão e Trabalhos Futuros

Este artigo apresentou um experimento preliminar, com o propósito de modelar a distribuição de uma espécie escolhida por meio do treinamento de um modelo. Este utilizou-se de dados integrados obtidos dos *datasets* do projeto *GOAmazon* juntamente

com datasets de ocorrência da espécie obtidos dos *datasets* do ICMBio e do GBIF. O *dataset* resultante da integração deste dados e que foi utilizado neste trabalho está disponibilizado em [Miyaji et al. 2021]. Destaca-se que, por mais que os resultados aqui apresentados tenham sido limitados a apenas uma das 40 espécies apresentadas na Tabela 1 pela questão da limitação do espaço, experimentos semelhantes foram realizados com as outras espécies, obtendo assim resultados parecidos relacionados ao desempenho do modelo.

O modelo possui bom desempenho, porém destaca-se a dificuldade de utilização dos dados em questão para o treinamento do modelo devido ao baixo volume destes, o que pode afetar sua acurácia e de se tratar de um sistema de classificação desbalanceada. Como trabalhos futuros pode-se reproduzir o mesmo experimento, porém utilizando dados provenientes dos voos realizados na estação chuvosa ou dados provenientes das estações de pesquisa, onde foram geradas séries temporais das variáveis atmosféricas. Além disso também é possível realizar experimentos com outros modelos clássicos que considerem a pseudo-ausência ou apenas dados de presença, como o modelo da Máxima Entropia.

## Agradecimentos

O presente trabalho foi possível devido a disponibilidade dos dados nos repositórios do *GOAmazon*, ICMBio e GBIF e ao Projeto Temático da FAPESP "Ciclos de vida e nuvens de aerossóis na Amazônia"(2017/ 17047-0).

## Referências

Berg, A., Gardenfors, U., and von Proschwitz, T. (2004). Logistic regression models for predicting occurrence of terrestrial molluscs in southern sweden: Importance of environmental data quality and model complexity. *Ecography*, 27(1):83–93.

Carneiro, L. R. d. A., Lima, A. P., Machado, R. B., and Magnusson, W. E. (2016). Limitations to the use of species-distribution models for environmental-impact assessments in the amazon. *PLoS One*, 11(1):e0146543.

Cirino, G., Brito, J., Barbosa, H. M., Rizzo, L. V., Tunved, P., de Sá, S. S., Jimenez, J. L., Palm, B. B., Carbone, S., Lavric, J. V., et al. (2018). Observations of manaus urban plume evolution and interaction with biogenic emissions in goamazon 2014/5. *Atmospheric Environment*, 191:513–524.

Effrosynidis, D., Tsikliras, A., Arampatzis, A., and Sylaios, G. (2020). Species distribution modelling via feature engineering and machine learning for pelagic fishes in the mediterranean sea. *Applied Sciences*, 10(24).

Elith, J. and Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *The Annual Review of Ecology, Evolution and Systematics*, 40:677–697.

Friedman, J. H. (2002). Stochastic gradient boosting. 38(4):367–378.

Hegel, T. M., Cushman, S. A., Evans, J., and Huettmann, F. (2010). *Current State of the Art for Statistical Modelling of Species Distributions*, pages 273–311. Springer Japan, Tokyo.

Hernandez, P. A., Graham, C. H., Master, L. L., and Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29(5):773–785.

Hutchinson, G. E. (1991). Population studies: Animal ecology and demography. *Bulletin of Mathematical Biology*, 53:193–213.

Johnson, R., Chawla, N., and Hellmann, J. (2012). Species distribution modeling and prediction: A class imbalance problem. pages 9–16.

Leidenberger, S., Obst, M., Kulawik, R., Stelzer, K., Heyer, K., Hardisty, A., and Bourlat, S. J. (2015). Evaluating the potential of ecological niche modelling as a component in marine non-indigenous species risk assessments. *Marine Pollution Bulletin*, 97(1):470–487.

Martin, S. T., Artaxo, P., Machado, L. A. T., Manzi, A. O., Souza, R. A. F. d., Schumacher, C., Wang, J., Andreae, M. O., Barbosa, H., Fan, J., et al. (2016). Introduction: observations and modeling of the green ocean amazon (goamazon2014/5). *Atmospheric Chemistry and Physics*, 16(8):4785–4797.

Mateo, R. G., Vanderpoorten, A., Muñoz, J., Laenen, B., and Désamoré, A. (2013). Modeling species distributions from heterogeneous data for the biogeographic regionalization of the european bryophyte flora. *PLoS One*, 8(2):e55648.

Miyaji, R. O., Almeida, F. V., Bueno, W., and Corrêa, P. L. P. (2021). Dados bioclimáticos para modelagem de distribuição de espécies baseada em variáveis ambientais e aerossóis na região próxima a manaus (am). Zenodo. DOI: 10.5281/zenodo.4824365.

Pinaya, J. (2019). Processo de modelagem paleoclimática de distribuição de espécies com enfoque nas mudanças climáticas. Tese apresentada à Escola Politécnica da Universidade de São Paulo.

Pinaya, J. and Corrêa, P. (2014). Metodologia para definição das atividades do processo de modelagem de distribuição de espécies. In *Anais do V Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais*, pages 45–54, Porto Alegre, RS, Brasil. SBC.

Radosavljevic, A. and Anderson, R. P. (2014). Making better maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, 41(4):629–643.

Scikit-learn Developers (2020). Scikit-learn: User guide. `https://scikit-learn.org/stable/user_guide.html`. Acesso em: 13/03/2021.

The Imbalanced-learn Developers (2021). Imbalanced-learn documentation. `https://imbalanced-learn.org/stable/`. Acesso em: 18/03/2021.

Wei, D., Fuentes, J. D., Gerken, T., Trowbridge, A. M., Stoy, P. C., and Chamecki, M. (2019). Influences of nitrogen oxides and isoprene on ozone-temperature relationships in the amazon rain forest. *Atmospheric Environment*, 206:280–292.

XGBoost Developers (2020). XGBoost Documentation. `https://xgboost.readthedocs.io/en/latest/`. Acesso em: 20/03/2021.