Utilização de dados espaciais para determinar a influência de poluentes na ocorrência de espécies na Amazônia

Renato O. Miyaji[1], **Pedro L. P. Corrêa**[1], **Luciana V. Rizzo**[2]

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

[2]Universidade Federal de São Paulo (UNIFESP)

{re.miyaji,pedro.correa}@usp.br

{lrizzo}@unifesp.br

Abstract. *The hydrological and energy cycles in the Amazon Basin region changed in the last decades, due to the influence of anthropic action. However, the effects of these on the local fauna have not yet been deeply analyzed. In this context, this work sought to develop an experiment of Species Distribution Modeling of birds, based on meteorological and aerosol data collected in the region of interest during the GoAmazon 2014/15 project, through the application of the Maximum Entropy Model, in order to determine the influence of pollutants on the occurrence of species.*

Resumo. *Por conta da influência da ação antrópica, os ciclos hidrológicos e energéticos na região da Bacia Amazônica sofreram alterações nas últimas décadas. No entanto, os efeitos dessas mudanças na fauna local ainda não foram profundamente analisados. Neste contexto, neste trabalho buscou-se desenvolver um experimento de Modelagem de Distribuição de Espécies de aves, a partir dos dados meteorológicos e de aerossóis coletados na região de interesse durante o projeto GoAmazon 2014/15, através da aplicação do Modelo de Máxima Entropia, de modo a determinar a influência de poluentes na ocorrência de espécies.*

1. Introdução

As condições climáticas na região central da Bacia Amazônica vêm sofrendo alterações nas últimas décadas, devido à expansão da região metropolitana de Manaus (AM). A população local está crescendo rapidamente, com taxas de aproximadamente 40 % a cada década desde 1960 [Martin et al. 2017]. Esse aumento da ação antrópica influencia as dinâmicas locais, como a distribuição de espécies da fauna. Porém, existe uma dificuldade para realizar análises com esse objetivo com uma maior profundidade, por conta da ausência de um grande volume de dados bioclimáticos confiáveis a respeito da região de interesse [Carneiro et al. 2016].

O projeto GoAmazon 2014/15, organizado pelo órgão *Atmospheric Radiation Measurement* (ARM) do Departamento de Energia dos Estados Unidos da América e por instituições nacionais, como a Universidade de São Paulo (USP), a Universidade do Estado do Amazonas (UEA) e o Instituto Nacional de Pesquisas Espaciais (INPE), buscou aumentar o volume de dados meteorológicos e de aerossóis disponíveis sobre a região localizada entre os municípios de Manaus (AM) e Manacapuru (AM). Isso foi feito por

meio de uma extensa coleta de dados entre os anos de 2014 e 2015, utilizando nove estações de coleta fixas e móveis, como aeronaves que realizaram voos de baixa altitude [Martin et al. 2016].

A partir desses dados, no trabalho de [Miyaji et al. 2021] foram aplicadas técnicas de interpolação espacial com o objetivo de realizar a inclusão de dados [Bertsimas et al. 2018] sobre as principais variáveis meteorológicas e de poluentes coletadas durante o projeto GoAmazon 2014/15. Assim, através da junção desse conjunto de dados com outro referente a ocorrência de espécies na mesma região, é possível viabilizar análises a respeito da biodiversidade local. Uma alternativa para essas análises é a Modelagem de Distribuição de Espécies. Essa técnica visa determinar a influência de variáveis ambientais na ocorrência de uma espécie e seu nicho ecológico - as condições que tornam um determinado habitat adequado para a sua ocorrência [Hutchinson 1991].

Assim, este trabalho buscou utilizar dados espaciais para desenvolver um experimento de Modelagem de Distribuição de Espécies com a aplicação do Modelo de Máxima Entropia [Phillips et al. 2004], para determinar a influência de poluentes na ocorrência de espécies de aves na região da Bacia Amazônica próxima às cidades de Manaus (AM) e Manacapuru (AM).

2. Trabalhos Relacionados

Experimentos de Modelagem de Distribuição de Espécies na região amazônica foram realizados anteriormente na literatura. No trabalho de [Carneiro et al. 2016], foi aplicado o Modelo de Máxima Entropia [Phillips 2005] para espécies de anfíbios, com base em variáveis de cunho geográfico e ambiental. Porém, notou-se a existência de grandes limitações quanto à disponibilidade de dados de distribuição para grande parte das espécies de interesse. Dessa forma, por conta do baixo volume de dados, os resultados dos modelos poderiam ser extrapolações de correlações espúrias, não possuindo significância estatística.

No experimento realizado por [Almeida et al. 2021], foram utilizados dados interpolados a partir dos coletados pelo projeto GoAmazon 2014/15, resultantes do trabalho de [Miyaji et al. 2021]. Assim, foram analisadas principalmente variáveis meteorológicas e de aerossóis na região da Bacia Amazônica. Para a tarefa da Modelagem de Distribuição de Espécies de pássaros, foram aplicados dois classificadores de Aprendizado de Máquina: a Regressão Logística e o algoritmo *Extreme Gradient Boosting* (XGBoost) [XGBoost Developers 2020]. Entretanto, para ambos o desempenho obtido foi negativamente afetado pela maior proporção de classes negativas (ausência da espécie) em relação às positivas (presença da espécie), configurando uma tarefa de classificação desbalanceada.

Esse resultado é recorrente quando se utiliza um conjunto de dados de presença-ausência [Hernandez et al. 2006], devido ao desbalanceamento que ocorre por conta da dificuldade em se afirmar a verdadeira ausência de uma espécie [Hegel et al. 2010]. Assim, outras técnicas para a construção do conjunto de dados utilizado podem aprimorar o desempenho dos modelos. Uma delas é dos modelos de pseudo-ausência (*pseudo-absence*), que consideram os dados de presença real em conjunto com uma amostra aleatória para representar os dados de ausência, além dos modelos de presença absoluta (*presence-only*), que levam em consideração apenas os registros de ocorrência da espécie

[Golini 2011], como o Modelo de Máxima Entropia.

Dessa forma, nota-se que a aplicação de diferentes tipos de modelos, como de pseudo-ausência e de presença absoluta, que utilizem dados meterológicos e de aerossóis para a Modelagem de Distribuição de Espécies podem ser úteis para aprimorar o desempenho das análises sobre a biodiversidade da região amazônica.

3. Metodologia

Para o desenvolvimento do trabalho, adotou-se a metodologia de Modelagem de Distribuição de Espécies com modelos estatísticos proposta por [Pinaya and Corrêa 2014], que pode ser vista na Figura 1. A hipótese científica adotada foi definida como a existência da influência da maior concentração de poluentes atmosféricos na ocorrência da espécie de ave a ser analisada.

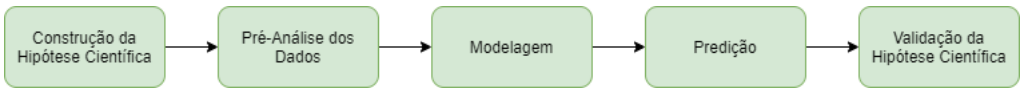

Figura 1. Metodologia adotada para experimento de Modelagem de Distribuição de Espécies [Pinaya and Corrêa 2014]

Na etapa de pré-análise dos dados, foi necessário obter o conjunto de dados bioclimáticos. As potenciais variáveis preditoras, ou seja, as meteorológicas e de aerossóis foram disponibilizadas por [Miyaji et al. 2021]. Esse cojunto de dados é resultante da aplicação da técnica de interpolação espacial linear segmentada sobre os dados coletados por uma das aeronaves do projeto GoAmazon 2014/15 durante 19 voos de baixa altitude realizados entre setembro e outubro de 2014.

O tratamento, a análise e a manipulação desse conjunto de dados foram realizados na linguagem *Python*, utilizando a aplicação *web Jupyter Notebook*. Dessa forma, foi possível obter as camadas de cada uma das variáveis, em função das coordenadas geográficas de latitude e longitude. A visualização construída com o uso da biblioteca *Matplotlib* da camada de concentração de monóxido de carbono (CO) em partes por milhão (ppm) é apresentada na Figura 2. Os dados obtidos compreendiam uma área de aproximadamente 6900 km^2 entre Manaus (AM) e Manacapuru (AM), com uma resolução espacial de 0,001° em latitude e longitude.

Para a construção do conjunto de dados bioclimáticos, também foi necessário coletar os dados de ocorrência de espécies. Esses foram obtidos de duas fontes: o Portal da Biodiversidade do Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio) e do *Global Biodiversity Information Facility* (GBIF), sendo posteriormente unidos, considerando as colunas de latitude, longitude, data do registro e espécie observada. Então, utilizando as funções da biblioteca *Pandas*, foi feito um filtro dos dados pela data, sendo considerados apenas os registros que correspondiam ao mesmo período dos dados climáticos. Em seguida, foi aplicada a operação de junção dos dois conjuntos de dados, considerando a chave composta de latitude e longitude.

Dessa forma, foi possível determinar as espécies com maiores quantidades de registros de presença no conjunto de dados bioclimáticos obtido. Foi escolhida a espécie *Tyrannus melancholicus*, o suiriri, que possuía 50 pontos distintos de ocorrência e pouca

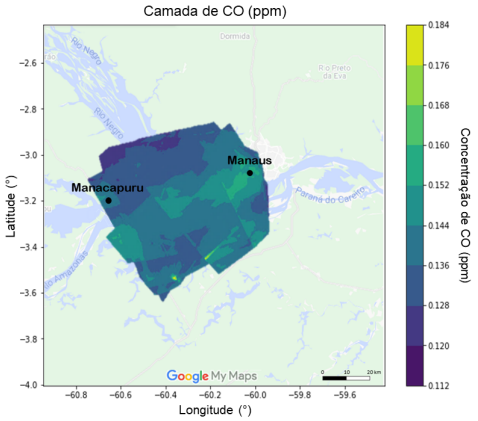

Figura 2. Camada de concentração de monóxido de carbono (CO) em partes por milhão (ppm)

dispersão espacial na região de interesse, fatores que contribuem para um melhor desempenho dos modelos [Hernandez et al. 2006].

Com a espécie de interesse definida, passou-se para a etapa de modelagem. Assim, foi feita uma análise da correlação entre as variáveis climáticas aos pares com o cálculo do coeficiente de correlação linear de Pearson [Mateo et al. 2013], de modo a se selecionar as variáveis preditoras. Para os pares que apresentavam esse coeficiente com módulo superior a 80 %, selecionou-se apenas uma das variáveis para ser utilizada, para evitar a ocorrência de multicolinearidade [Mateo et al. 2013]. Desse modo, as variáveis selecionadas foram: as temperaturas máxima e mínima, as concentrações de monóxido de carbono (CO), ozônio (O_3), óxidos de nitrogênio (NO_X), metano (CH_4), isopreno (C_5H_8), acetonitrila (CH_3CN) e a fração volumétrica de água (H_2O). Assim, o conjunto de dados bioclimáticos a ser utilizado possuía nove atributos e uma variável-resposta: a ocorrência de *Tyrannus melancholicus*.

Selecionou-se o Modelo de Máxima Entropia para ser aplicado, uma vez que esse apresenta os melhores desempenhos quando o conjunto de dados que se dispõe é de média ou baixa dimensionalidade [Phillips 2005]. Como se trata de um modelo de presença absoluta, foram considerados somente os registros distintos de presença da espécie. O método baseia-se na analogia existente com a termodinâmica. Assim como para a entropia física, a interpretação da entropia da informação seria da medida do desconhecimento do sistema em análise. Dessa forma, quanto maior é essa, menos conhecimento se tem a respeito do processo. Portanto, maximizando o seu valor, é possível encontrar a distribuição de probabilidade que é estatisticamente mais provável de ser verdadeira [Phillips 2005]. A aplicação do modelo foi feita através do pacote *maxnet* da linguagem R, desenvolvido por [Phillips et al. 2017].

Após o treinamento do modelo, prosseguiu-se para a etapa de predição. Nela, o desempenho do modelo foi avaliado em função da área sobre a curva característica de operação do receptor (*Area Under the Receiver Operating characteristic Curve* - AUC-ROC). Essa avalia a probabilidade que as classes sejam corretamente classificadas pelo modelo com a variação do limite de classificação. Essa métrica varia entre 0 e 1, sendo o valor unitário o ideal [Pinaya and Corrêa 2014]. Além disso, foi realizada a previsão para a área completa de dados climáticos disponível, gerando um mapa de distribuição

potencial, no qual para cada par de coordenadas de latitude e longitude conhecidas, o modelo atribui um valor entre 0 e 1, que representa a probabilidade de ocorrência da espécie naquele ponto analisado [Pinaya and Corrêa 2014].

Por fim, foi possível validar a hipótese científica. Isso foi feito a partir da determinação das curvas de resposta. Essas representam a dependência da probabilidade de ocorrência da espécie em função dos valores assumidos pelas variáveis climáticas [Phillips et al. 2017].

4. Resultados e Discussões

O Modelo de Máxima Entropia foi treinado sobre o conjunto de dados bioclimáticos obtido para a espécie *Tyrannus melancholicus*. Assim, foi possível determinar o AUC-ROC, obtendo um valor de 83 %. Nota-se que esse é elevado, indicando que o modelo possui uma boa capacidade preditora. Ademais, a partir do modelo treinado foi feita a previsão para toda a área de abrangência das camadas de variáveis meteorológicas e de aerossóis, obtendo o mapa de distribuição potencial da Figura 3.

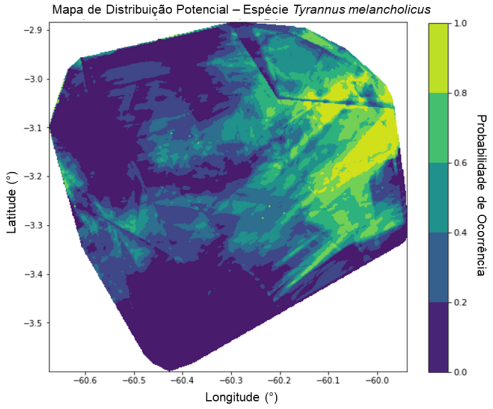

Figura 3. Mapa de distribuição potencial para espécie *Tyrannus melancholicus*

Com o objetivo de se validar a hipótese científica estabelecida, foram determinadas as curvas de resposta que podem ser vistas na Figura 4. Assim, nota-se que a espécie *Tyrannus melancholicus* possui alta sensibilidade a maior parte das variáveis preditoras consideradas. Especificamente a respeito de poluentes atmosféricos, observa-se uma redução da probabilidade de ocorrência da espécie em função do aumento das concetrações CO, O_3, NO_X, CH_4 e C_5H_8. Essa tendência mostrou-se mais acentuada para concentrações acima de 35 ppb de ozônio e de 1,79 ppm de metano, indicando uma maior sensibilidade da espécie para esses poluentes. Já para a acetonitrila, percebe-se um pequeno crescimento da probabilidade de ocorrência com o aumento da concentração desse poluente.

5. Conclusão e Trabalhos Futuros

A partir dos resultados obtidos, conclui-se que foi possível utilizar dados espaciais para desenvolver um experimento preliminar de Modelagem de Distribuição de Espécies para a ave *Tyrannus melancholicus*, na região da Bacia Amazônica próxima a Manaus (AM) e Manacapuru (AM). Com a utilização do Modelo de Máxima Entropia, pôde-se gerar

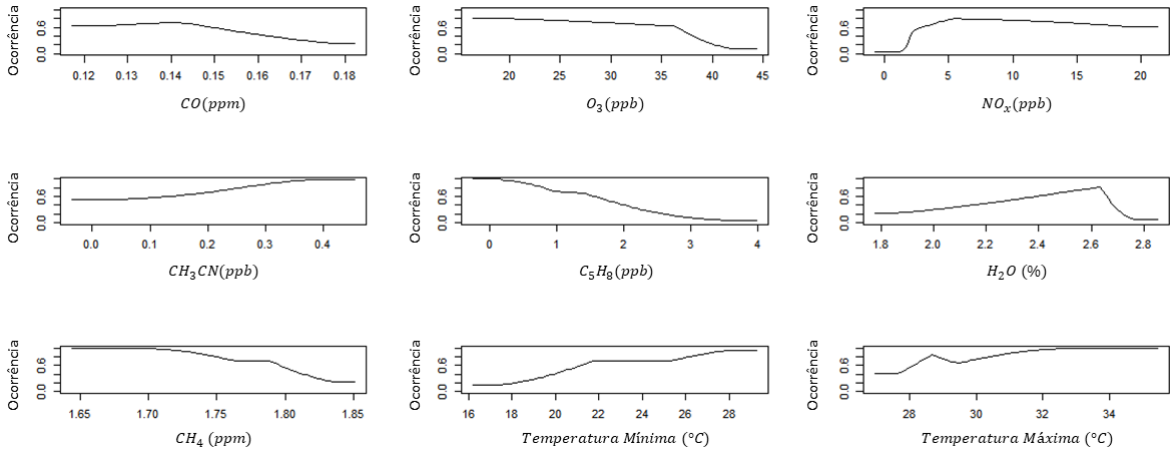

Figura 4. Curvas de resposta para espécie *Tyrannus melancholicus*

um mapa de distribuição potencial, que permitiu avaliar a probabilidade de ocorrência da espécie de estudo sobre a área analisada, além de gerar as curvas de resposta em função de cada variável preditora. Assim, a hipótese científica inicialmente estabelecida foi avaliada, concluindo que a espécie de estudo era especialmente sensível aos poluentes metano e isopreno.

Como a métrica AUC-ROC possui o valor ideal unitário, para futuros trabalhos, recomenda-se o aprimoramento do modelo. Isso pode ser feito com a incorporação de outras variáveis preditoras. Outra alternativa seria a aplicação de outras técnicas, como as de classificação de Aprendizado de Máquina, sobre um conjunto de dados de pseudo-ausência. Além disso, o modelo obtido é restrito à espécie *Tyrannus melancholicus*. Portanto, para uma melhor avaliação do Modelo de Máxima Entropia, esse também deveria ser aplicado para outras espécies.

Agradecimentos

Este trabalho foi possível devido ao apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), através de uma bolsa do programa PIBIC (2020/21 - 1745), dos Projetos Temáticos da FAPESP ”Ciclos de vida e nuvens de aerossóis na Amazônia”(2017/ 17047-0) e ”Research Centre for Greenhouse Gas Innovation - RCG2I”(2020/15230-5) e dos pesquisadores do Grupo de Pesquisa em Big Data e Ciência dos Dados da EPUSP.

Referências

Almeida, F. V., Bueno, W. M., Miyaji, R. O., and Corrêa, P. L. P. (2021). Experimento de modelagem de distribuição de espécies baseada em variáveis ambientais e de aerossóis na região próxima a manaus (am). In *Anais do XII Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*. SBC.

Bertsimas, D., Pawlowski, C., and Zhuo, Y. D. (2018). From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*, 18:1–39.

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.

Golini, N. (2011). *Bayesian Modelling of Presence-only Data*. PhD thesis, Spienza Universidade de Roma.

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(1-2):193–213.

Martin, S. T., Artaxo, P., Machado, L., Manzi, A. O., Souza, R. A. F. d., Schumacher, C., Wang, J., Biscaro, T., Brito, J., Calheiros, A., et al. (2017). 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*, 98(5):981–997.

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., Bauer, L. O., Ferrari, V. M., Almeida, F. V., Corrêa, P. L. P., and Rizzo, L. V. (2021). Interpolação espacial de variáveis ambientais e aerossóis na região da bacia amazônica próxima a manaus-am. In *Anais do XII Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*. SBC.

Phillips, S. J. (2005). Maximum entropy modeling of species geographic distribution. *Ecological Modelling*, 190:231–259.

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., and Blair, M. E. (2017). Opening the black box: an open-source release of maxent. *Ecography*, 40:887—-893.

Phillips, S. J., Dudik, M., and Schapire, R. E. A. (2004). Maximum entropy approach to species distribution modelling. *In Proceedings of the Twenty-First International Conference on Machine Learning*, pages 655–662.

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.

XGBoost Developers (2020). XGBoost Documentation. `https://xgboost.readthedocs.io/en/latest/`. Acesso em: 01/07/2021.