Handling uncertainty through Bayesian inference for Species Distribution Modelling in the Amazon Basin region

Renato O. Miyaji[1], **Pedro L. P. Corrêa**[1]

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

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

Abstract. *Species Distribution Modelling (SDM) is used for biodiversity preservation and wildlife management. In order to apply it, a large and reliable dataset about the species occurrence is required. However, in some cases, this can be difficult when there are only a few occurrence records. In this context, uncertainty handling techniques can be applied. Thus, Bayesian inference was used in this study to perform SDM in the Amazon Basin region near Manaus (AM) with data collected by the GoAmazon 2014/15 project. The results were compared with those obtained in statistical models and were similar.*

Resumo. *Uma das ferramentas mais utilizadas para o monitoramento da biodiversidade é a modelagem de distribuição de espécies. Para a sua aplicação, é necessário possuir uma grande base de dados confiáveis a respeito da ocorrência de espécies. Entretanto, essa condição não é satisfeita quando existem poucos registros de ocorrência. Nesse contexto, podem ser aplicadas técnicas de tratamento de incertezas. Assim, este trabalho buscou utilizar a abordagem Bayesiana para permitir a modelagem de distribuição de espécies na região da Bacia Amazônica próxima a Manaus (AM), com base em dados coletados pelo projeto GoAmazon 2014/15. Os resultados foram comparados com os resultantes de técnicas clássicas, obtendo desempenhos semelhantes.*

1. Introdução

A Floresta Amazônica é uma região de grande interesse, por conta de sua biodiversidade. Localizada no centro da floresta, a cidade de Manaus (AM) apresenta-se como um laboratório ideal para estudar a influência da ação antrópica no clima e nos ecossistemas terrestres em uma floresta tropical [Martin et al. 2017]. Para isso, entre 2014 e 2015 foi desenvolvido o projeto GoAmazon 2014/15 pelo *Atmospheric Radiation Measurement* (ARM), vinculado aos Estados Unidos da América, em conjunto com instituições brasileiras. A partir dele, foram coletados dados meteorológicos e de aerossóis, por meio de voos de baixa altitude.

O Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio) monitora a biodiversidade nacional, disponibilizando dados de ocorrência de espécies sobre todo o território do país através do Portal da Biodiversidade [ICMBio 2021]. No entanto, para a região próxima à cidade de Manaus (AM), dispõe-se de uma baixa quantidade de dados para uma mesma espécie de pássaros, o que dificulta a aplicação de técnicas estatísticas para a análise da biodiversidade local [Almeida et al. 2021].

O efeito de variáveis ambientais na ocorrência de espécies pode ser analisado por meio da aplicação de modelos de distribuição de espécies (*Species Distribution Models*

– SDM). Esses permitem que se avalie o nicho ecológico para a espécie de interesse, isto é, as condições ambientais que tornam um habitat adequado para a sua ocorrência [Hutchinson 1991].

Com os grandes avanços ocorridos nas últimas décadas na área de Aprendizado de Máquina, foram desenvolvidos modelos com desempenhos cada vez melhores. Dessa forma, a utilização desses tornou-se mais frequente para a Modelagem de Distribuição de Espécies [Hegel et al. 2010]. Para a aplicação desses, uma condição necessária é a disponibilidade de uma grande base de dados confiáveis a respeito da ocorrência de espécies. Entretanto, essa não é satisfeita em diversos contextos, principalmente quando se tratam de dados relacionados a espécies raras, que podem estar em extinção ou localizadas em áreas de difícil acesso. Nesses casos, o uso de uma pequena quantidade de dados ou de dados não confiáveis, pode levar a resultados errados [Martin et al. 2005].

Como alternativa, pode-se recorrer a técnicas de tratamento de incertezas, como a inferência Bayesiana e a teoria da evidência de Dempster-Shafer. Para isso, pode-se utilizar opiniões de especialistas, resultados de experimentos prévios ou meta análises como conhecimento prévio para o modelo selecionado [Low-Choy et al. 2009].

Assim, este trabalho buscou avaliar se a aplicação de técnicas de tratamento de incertezas poderia viabilizar a modelagem de distribuição de espécies de pássaros com poucos registros de ocorrência na região da Bacia Amazônica próxima à cidade de Manaus (AM), utilizando os dados climáticos coletados pelo projeto GoAmazon 2014/15.

2. Trabalhos Relacionados

Alternativas à abordagem estatística para a modelagem de distribuição de espécies quando os dados disponíveis são poucos ou não são confiáveis foram apresentadas na literatura. [Ellison 2004] realizou um levantamento bibliográfico do uso da abordagem Bayesiana para a modelagem de distribuição de espécies. Através dele, o autor identificou 69 artigos, sendo que a maioria desses fez o uso de informações de estudos prévios como forma de determinar a distribuição de probabilidade *a priori*. O restante aplicou uma metodologia de elicitação para a obtenção de conhecimentos de especialistas.

[Kunhert et al. 2010, Low-Choy et al. 2009, Martin et al. 2005] o fizeram a partir da elicitação da opinião de especialistas, ou seja, pesquisadores que acumularam conhecimento a respeito de um tema de interesse através de experiências vividas, treinamento e aprendizado. Posteriormente, foi aplicado diretamente o Teorema de Bayes.

Uma outra maneira de se utilizar a abordagem Bayesiana é através de modelos Bayesianos. Neles, é possível fornecer as distribuições de probabilidade *a priori* e de verossimilhança (*likelihood*) a respeito de seus parâmetros. Então, pode-se determinar a distribuição de probabilidade *a posteriori* deles. A regressão logística é um classificador amplamente utilizado para essa tarefa [Hegel et al. 2010]. Esse, em sua versão Bayesiana, foi avaliado por [Di Lorenzo et al. 2011] e [Golini 2011], concluindo que se trata de uma opção relevante para o tratamento de incertezas.

Outra alternativa possível é a aplicação da Teoria da Evidência de Dempster-Shafer. [Niamir 2019] fez o uso dessa a partir de conhecimentos de especialistas a respeito das interações entre ambiente e espécies, considerando e incorporando as incertezas do processo.

3. Metodologia

3.1. Definição do Modelo

A partir da revisão da literatura realizada, notou-se que a principal técnica utilizada no contexto de modelagem de distribuição de espécies foi a Bayesiana. Para isso, podem ser adotadas diversas fontes para o fornecimento das probabilidades *a priori* e de verossimilhança. Como a precisão da abordagem através da incorporação de conhecimentos advindos de especialistas possui uma dependência forte com a quantidade de especialistas consultados [Kunhert et al. 2010], neste trabalho optou-se pelo uso de modelos Bayesianos.

O modelo utilizado foi o de regressão logística Bayesiana, por apresentar resultados promissores para conjuntos de dados de baixa e média dimensão [Golini 2011]. Sua escolha foi feita de modo a possibilitar o tratamento de incertezas envolvendo os dados de ocorrência de espécies. Essas podem ser decorrentes de erros humanos, como a não identificação da espécie em uma determinada região no momento exato. Assim, enquanto a ocorrência de uma espécie pode ser realmente determinada, a afirmação de sua ausência é uma tarefa difícil [Hegel et al. 2010].

Nesse sentido, existem três classes de modelos possíveis. A primeira delas é dos classificadores tradicionais de aprendizado de máquina que consideram os dados de presença e ausência. Para eles, além das incertezas citadas anteriormente devido à afirmação da ausência, sua acurácia é prejudicada, por se tratar de uma tarefa de classificação desbalanceada, na qual a proporção de classes negativas (ausências) é muito maior que a de positivas (presença) [Johnson et al. 2012]. A segunda classe é dos modelos que consideram apenas os dados de presença real (*presence-only*) e, a partir deles, sumarizam as características de adequabilidade e as extrapolam [Golini 2011]. Um dos modelos mais utilizados pertencentes a essa classe é o de Máxima Entropia, porém esse não permite o tratamento de incertezas [Phillips 2005]. Por fim, a terceira classe é dos modelos de pseudo-ausência (*pseudo-absence*). O conjunto de dados utilizado por esses é composto pelos registros nos quais foi observada a ocorrência da espécie e por outros registros amostrados que representam os pontos de ausência da espécie. Para tal, recomenda-se o uso de uma amostragem aleatória sem reposição [Golini 2011].

Assim, selecionou-se o modelo de regressão logística Bayesiana de pseudo-ausência para ser aplicado neste trabalho, devido aos bons desempenhos obtidos por [Di Lorenzo et al. 2011] e [Golini 2011].

Desse modo, foi realizada a modelagem de distribuição de espécies de pássaros na região da Bacia Amazônica entre as cidades de Manaus (AM) e Manacapuru (AM), considerando a influência de variáveis meteorológicas e de aerossóis na ocorrência dessas. Os dados climáticos utilizados são provenientes de interpolações espaciais realizadas a partir de coletas feitas na região por aeronaves do projeto GoAmazon 2014/15 [Miyaji et al. 2021]. A temperatura, as concentrações de ozônio (O_3), monóxido de carbono (CO), óxidos de nitrogênio (NO_X), metano (CH_4), dióxido de carbono (CO_2), isopreno e acetonitrila, a concentração numérica de partículas (CPC 3010) e a fração volumétrica de água (H_2O) foram as variáveis disponibilizadas.

A fim de se possibilitar a comparação dos resultados obtidos por meio da abordagem Bayesiana com os resultantes de métodos estatísticos, foi selecionada uma espécie

pertencente à classe *Aves* que possuísse registros de ocorrência na região durante os anos de 2014 e 2015, disponibilizados pelo Instituto Chico Mendes de Conservação da Biodiversidade através do Portal da Biodiversidade [ICMBio 2021]. A técnica estatística utilizada como base de comparação foi o modelo de Máxima Entropia, por geralmente apresentar os melhores resultados em relação aos demais para bases de dados pequenas e médias [Phillips 2005].

3.2. Coleta e Tratamento de Dados

De modo a se construir o conjunto de dados necessário para a aplicação da modelagem de distribuição de espécies, coletou-se as variáveis meteorológicas e de aerossóis interpoladas na região entre Manaus (AM) e Manacapuru (AM) de [Miyaji et al. 2021]. Em seguida, foi realizada a coleta dos dados de ocorrência de espécies a partir do Portal da Biodiversidade [ICMBio 2021].

Então, utilizando a linguagem *Python*, através da aplicação *web Jupyter Notebook*, aplicou-se um filtro, selecionando os registros durante o mesmo período e na mesma localização do projeto GoAmazon 2014/15. Desse modo, obteve-se registros de ocorrência de 40 espécies diferentes, 39 da classe *Aves* e uma da *Reptilia*. Foi definido outro filtro: a espécie deveria apresentar a partir de 17 ocorrências para ser considerada [Pinaya 2019]. Dentre as espécies com a maior quantidade de ocorrência, destacaram-se a *Coragyps atratus* e a *Tyrannus melancholicus*, com 54 e 50 ocorrências distintas, respectivamente.

Assim, foi possível construir um conjunto de dados bioclimáticos, por meio da operação de junção dos dados ambientais e de aerossóis com os de ocorrência de espécies, considerando a latitude, a longitude e a data de registro da ocorrência.

A fim de se determinar os atributos que seriam considerados pelo modelo de distribuição de espécies, realizou-se uma análise a partir da matriz de correlação, apresentada na Figura 1. Utilizou-se como métrica o coeficiente de Pearson. Esse é capaz de analisar aos pares a relação linear entre os atributos do cojunto de dados. Retirou-se um dos atributos dos pares que apresentavam o coeficiente de Pearson com módulo a partir de 80 % [Mateo et al. 2013], ou seja, com alta correlação linear entre si. Dessa forma, buscou-se que o modelo não incorporasse padrões aleatórios e que não ocorresse o fenômeno de multicolinearidade [Pinaya and Corrêa 2014]. Assim, retirou-se três atributos: temperatura, concentração de dióxido de carbono e concentração numérica de partículas (CPC).

3.3. Treinamento dos Modelos

Para o estudo de caso, foi selecionada a espécie com a maior quantidade de observações reais de ocorrência, porém que ainda caracterizava um desbalanceamento no conjunto de dados: a *Coragyps atratus*, o urubu-de-cabeça-preta. O conjunto de dados contendo as classes positivas para essa espécie foi considerado. Então, foi amostrado aleatoriamente dos demais dados um conjunto de mesma dimensão, de modo a representar as amostras de ausência da espécie. Assim, concatenando os dois conjuntos, obteve-se o final que poderia ser utilizado para a modelagem de distribuição de espécies. Esse conjunto de dados bioclimáticos possuía nove atributos e uma variável resposta: a ocorrência de *Coragyps atratus*.

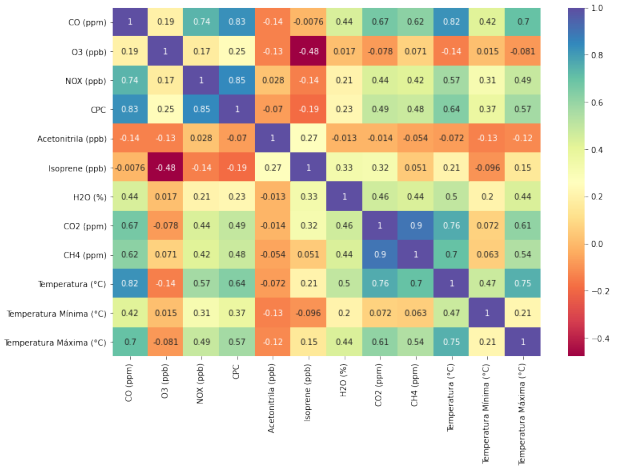

Figura 1. Matriz de correlação entre os atributos

Então, foram definidos os modelos que seriam avaliados e comparados. Todos os modelos seriam de regressão logística Bayesiana, na qual, diferentemente da abordagem frequentista, pode-se informar as distribuições de probabilidade *a priori* para cada parâmetro do modelo e posteriormente, após o cálculo da distribuição de probabilidade de verossimilhança, utilizar o Teorema de Bayes para determinar a distribuição *a posteriori* para cada um deles [Di Lorenzo et al. 2011]. A regressão logística se trata de um classificador linear definido a partir da equação (1). Nela, os parâmetros a serem ajustados são β_i com $i = 0, \ldots, n$, sendo n a quantidade de atributos considerados.

$$P(Y = 1/X) = \frac{e^{\eta(X)}}{1 + e^{\eta(X)}} \Rightarrow logit(P(Y = 1/X)) = \sum_{i=i}^{n} \beta_i X_i \tag{1}$$

A determinação da distribuição de probabilidade *a posteriori* desses parâmetros é feita por meio do Teorema de Bayes, de acordo com a equação (2), na qual D são os dados observados e θ é o parâmetro da distribuição. No entanto, essa pode ser uma tarefa difícil. Como alternativa, pode-se utilizar a técnica Cadeias de Markov Monte Carlo (MCMC). Essa gera Cadeias de Markov que devem convergir na distribuição de probabilidade *a posteriori* de interesse a partir da simulação de Monte Carlo. O algoritmo utilizado para tal é composto de algumas etapas: a definição de valores iniciais dos parâmetros desconhecidos, de modo a representar a variância dos componentes. Em seguida, a amostragem de cada parâmetro, enquanto os demais são mantidos constantes. Por fim, os valores amostrados de cada parâmetro são monitorados até a convergência [Martin et al. 2005]. A construção dos modelos utilizando a técnica MCMC foi feita na linguagem *Python* através da biblioteca *PyMC3* [Salvatier et al. 2016].

$$P(\theta/D) \propto P(D/\theta)P(\theta) \tag{2}$$

As variantes do modelo de regressão logística Bayesiana a serem comparadas foram: um modelo utilizando os nove atributos, porém com distribuições de probabilidades *a priori* não informativas para os parâmetros, sendo essas distribuições normais com média zero e variância igual a 100 ($\beta_i \sim N(0, 100)$), de modo a permitir o ajuste em uma

faixa ampla. As distribuições de probabilidade adotadas para o primeiro modelo podem ser vistas na Figura 2 a). O segundo modelo avaliado seria utilizando distribuições de probabilidade *a priori* informativas, a partir de análises prévias realizadas, como sugerido por [Low-Choy et al. 2009].

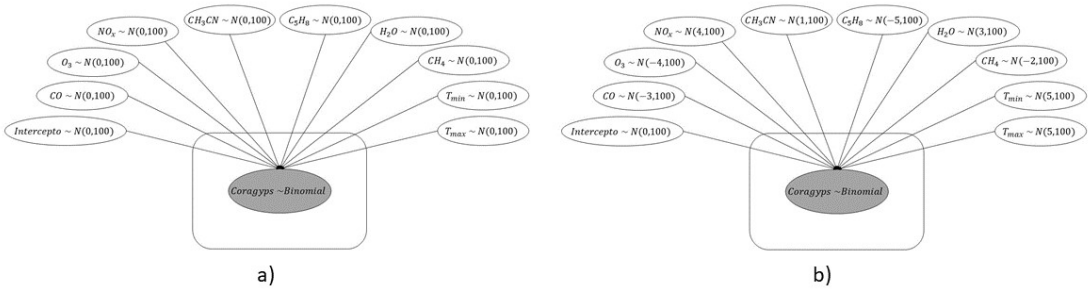

Figura 2. Modelo de regressão logística Bayesiana com distribuição de probabilidades a priori a) não informativas; b) informativas

Para determinar as distribuições de probabilidade *a priori* informativas, foram utilizados os resultados obtidos através da modelagem de distribuição de espécies para *Coragyps atratus* pelo Modelo de Máxima Entropia. Através desse, observou-se a curva de resposta da ocorrência da espécie em função de cada atributo considerado, como apresentado na Figura 3. Dessa forma, a fim de se determinar a distribuição *a priori* de cada atributo, o sinal da média da distribuição normal foi adotado como sendo o mesmo da tendência global observada para a curva de resposta com o aumento do atributo considerado. Já a magnitude da média foi proporcional ao módulo da variação observada para a probabilidade de ocorrência na curva de resposta. Assim, as distribuições de probabilidade *a priori* para o modelo informativo podem ser vistas na Figura 2 b).

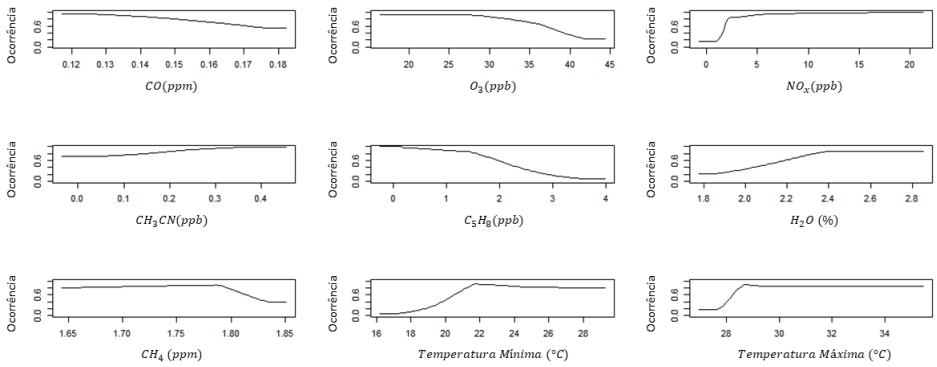

Figura 3. Curvas de resposta para a espécie *Coragyps atratus* obtidas pelo Modelo de Máxima Entropia

Para cada um dos modelos, avaliou-se a convergência das distribuições *a posteriori* dos parâmetros. Também foram analisadas as distribuições para cada um deles, assim como seu intervalo de alta densidade (*High Density Interval* – HDI) [Salvatier et al. 2016]. Então, para validá-los, foi utilizada a técnica de verificação de preditiva posterior (*Posterior Predictive Check* - PPC). Através dessa, são amostrados valores de cada distribuição *a posteriori* e esses são utilizados para gerar pre-

visões. Essas são comparadas com os valores observados, de modo a avaliar o modelo [Gelman and Rubin 1995].

Então, podem ser avaliadas as métricas relacionadas à classificação, como a acurácia, a precisão e a revocação. Além disso, pode-se utilizar o teste AUC-ROC (*Area Under the Receiver Operating characteristic Curve*). A interpretação desse é a probabilidade que instâncias positivas e negativas sejam corretamente classificadas pelo algoritmo. O parâmetro AUC varia de 0 a 1, sendo o valor ideal o unitário [Phillips 2005].

Adicionalmente, como forma complementar de avaliação, o conjunto de dados foi dividido em partes de treinamento e de teste na proporção 70/30. Assim, para cada modelo foram determinados os parâmetros de máxima verossimilhança, sendo esses utilizados para realizar a previsão sobre o conjunto de dados de teste, calculando as métricas de acurácia, precisão, revocação e AUC-ROC.

4. Resultados e Discussões

As duas variantes do modelo de regressão logística Bayesiana foram ajustadas ao conjunto de dados bioclimáticos, construído com o uso do conceito de pseudo-ausência. Através da técnica MCMC, foi possível determinar as distribuições de probabilidade *a posteriori* para cada um dos parâmetros do modelo. Para avaliar a convergência do algoritmo MCMC, foi avaliada a métrica $\hat{R}$, como essa possuiu módulo inferior a 1,05 para todos os parâmetros, conclui-se que o houve convergência [Gelman and Rubin 1992]. Na Figura 4, são apresentadas as distribuições de probabilidade *a posteriori* para cada parâmetro do modelo com distribuições de probabilidade *a priori* não informativas. Já na Figura 5, são apresentadas as distribuições para o modelo com as com distribuições de probabilidade *a priori* informativas.

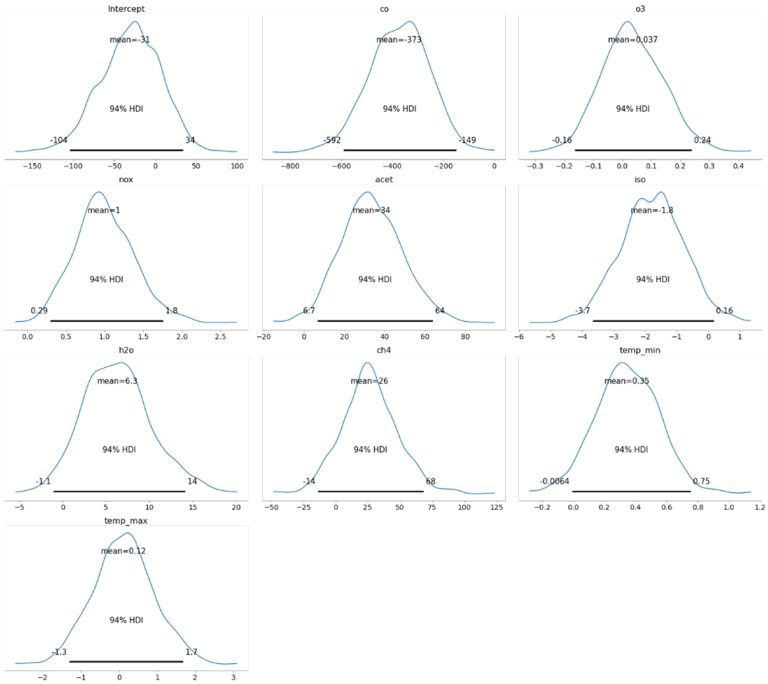

Figura 4. Distribuições de probabilidade a posteriori para cada parâmetro do modelo com distribuições de probabilidade a priori não informativas

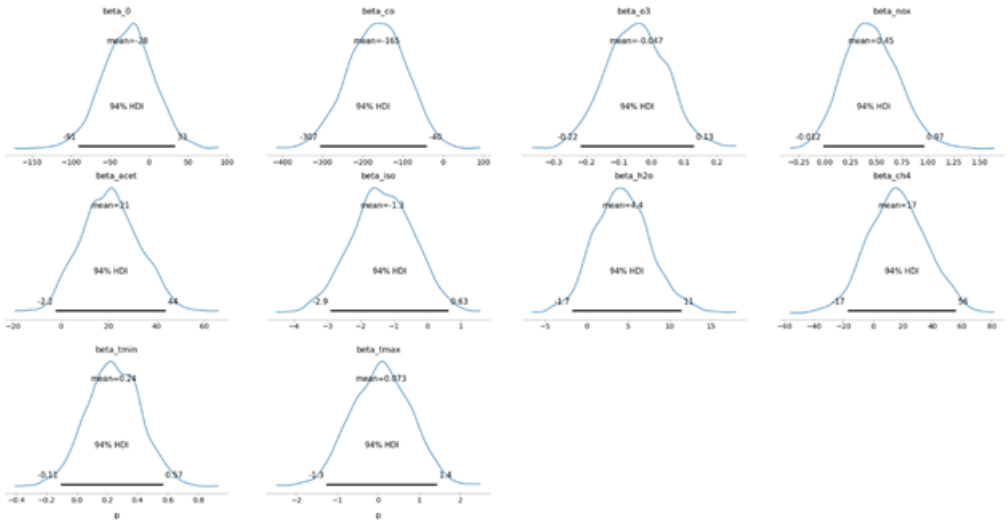

Figura 5. Distribuições de probabilidade a posteriori para cada parâmetro do modelo com distribuições de probabilidade a priori informativas

A partir das Figuras 4 e 5, também é possível observar o HDI, assim como a média das distribuições. Para alguns parâmetros, como os relacionados com as variáveis concentração de CO, NO_X, CH_4 e temperatura máxima, nota-se que houve uma variação mais expressiva na média das distribuições de probabilidade para o modelo com distribuições informativas, em comparação com o usando distribuições não informativas. Para os demais, a média das distribuições permaneceu próxima. Considerando o modelo com distribuições de probabilidade *a priori* informativas, percebe-se que houve uma variação significativa no módulo das médias em comparação com as probabilidades *a priori*. Porém, para todas, exceto CH_4, o sinal manteve-se o mesmo. Esse comportamento era esperado, por conta do desvio padrão de 100 que foi adotado, que permitia liberdade para o ajuste dos parâmetros.

Para cada um dos modelos, foi realizada uma validação através do emprego da técnica de verificação de preditiva posterior (PPC). Dessa forma, foi possível comparar a previsão do modelo com o valor observado. Assim, calculou-se as métricas de classificação de acurácia, precisão e revocação para cada modelo, além do AUC ROC, como apresentado na Tabela 1.

	Modelo Não Informativo	Modelo Informativo
Acurácia	72 %	72 %
Precisão	69 %	71%
Revocação	67 %	61%
AUC ROC	81 %	79 %

Tabela 1. Métricas de classificação a partir da verificação preditiva posterior

A partir da análise da Tabela 1, nota-se que o desempenho de classificação de ambos os modelos foi semelhante, com acurácia igual. O modelo com distribuições de probabilidade *a priori* informativas apresentou uma precisão de dois pontos percentuais

maior em relação ao outro. Por outro lado, o modelo com distribuições de probabilidade *a priori* não informativas obteve um AUC ROC de dois pontos percentuais maior e uma revocação de seis pontos percentuais mais elevada. Assim, a partir da verificação de preditiva posterior, o modelo com distribuições de probabilidade *a priori* não informativas mostrou-se superior.

Como forma de validação adicional, foi comparada a previsão de cada modelo para os dados do conjunto de teste. Para isso, foi necessário obter as estimativas de máxima verossimilhança de cada parâmetro dos modelos. Então, foram calculadas novamente as métricas de classificação, apresentadas na Tabela 2.

	Modelo Não Informativo	Modelo Informativo
Acurácia	53 %	59 %
Precisão	71 %	79%
Revocação	48 %	52%
AUC ROC	66 %	62 %

Tabela 2. Métricas de classificação a partir da previsão no cojunto de teste

Em comparação com os resultados obtidos através da verificação de preditiva posterior, percebe-se que, para os dados do conjunto de teste, o desempenho de ambos os modelos foi inferior. Novamente, não houve uma diferença muito grande no desempenho entre os modelos. Porém, neste caso, o modelo com distribuições de probabilidade *a priori* informativas apresentou melhores resultados em termos de acurácia, precisão e revocação. Apenas para o AUC ROC que o desempenho do modelo não informativo foi superior.

Então, foi feita uma previsão com cada modelo para a área completa de dados climáticos disponível, de modo a se construir mapas de distribuição potencial, apresentados na Figura 6. Com o Modelo de Máxima Entropia, foi possível obter o mapa de distribuição potencial também apresentado na Figura 6.

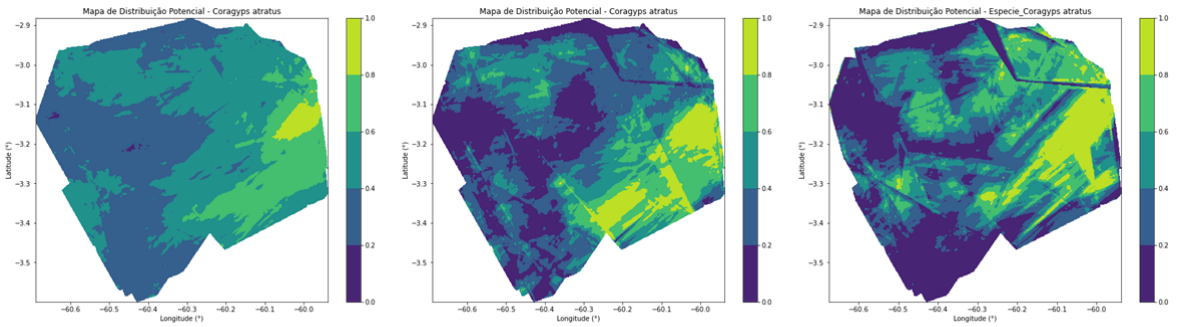

Figura 6. Mapa de distribuição potencial de *Coragyps atratus* a partir de a) modelo com distribuições de probabilidade a priori não informativas; b) modelo com distribuições de probabilidade a priori informativas; c) modelo de Máxima Entropia

A partir da análise da Figura 6, nota-se que o mapa de distribuição potencial obtido através do modelo com distribuições de probabilidade *a priori* informativas apresen-

tou uma semelhança maior com o obtido pelo Modelo de Máxima Entropia, tanto em termos de valores absolutos, quanto em relação à variabilidade espacial. O mapa obtido pelo modelo com distribuições de probabilidade *a priori* não informativas apresentou uma maior suavidade das curvas, com regiões maiores com uma mesma probabilidade média de ocorrência. De forma geral, ambos os modelos testados apresentaram a mesma tendência: com maiores valores de probabilidade de ocorrência da espécie na região leste do mapa, nas proximidades da cidade de Manaus (AM), onde se concentravam as maiores ocorrências reais da espécie *Coragyps atratus*, e com probabilidades menores na região oeste próxima ao munícipio de Manacapuru, onde existiam poucos pontos de observação da espécie.

Assim, apesar do desempenho obtido pelos modelos por meio da verificação de preditiva posterior e da previsão no conjunto de dados de teste ter sido semelhante em relação às métricas de classificação, o mapa de distribuição potencial obtido pelo modelo com distribuições de probabilidade *a priori* informativas apresentou maior similaridade com o obtido pelo Modelo de Máxima Entropia. Portanto, conclui-se que o uso de distribuições de probabilidade *a priori* informativas, mesmo que por meio de estudos prévios, possui uma grande importância para a modelagem de distribuição de espécies.

5. Conclusão e Trabalhos Futuros

Neste trabalho, foi possível avaliar a viabilidade do uso de técnicas de tratamento de incertezas para a tarefa de modelagem de distribuição de espécies de pássaros na região da Bacia Amazônica. Para isso, foi realizada uma revisão bibliográfica, de modo a determinar as possíveis abordagens que poderiam ser adotadas. Devido à dificuldade da realização de uma grande quantidade de pesquisas com especialistas do domínio, optou-se pelo uso de informações de estudos prévios como distribuições de probabilidade *a priori*. Por conta de resultados promissores relatados na literatura, selecionou-se um modelo de regressão logística Bayesiana. Para esse, foi construído um conjunto de dados a partir do conceito de pseudo-ausência. Para a seleção dos atributos, foi feita uma análise da matriz de correlação, determinando um conjunto de nove variáveis.

Então, foram definidos dois modelos para comparação: um contendo distribuições de probabilidade *a priori* não informativas e outro com distribuições de probabilidade *a priori* informativas, a partir de estudos prévios. Através do algoritmo de Cadeias de Markov com simulação de Monte Carlo (MCMC), foi possível obter as distribuições de probabilidade *a posteriori* para cada parâmetro dos modelos. A fim de se validar os modelos, foi realizada uma verificação de preditiva posterior. Além disso, para avaliar o desempenho do modelo frente a novos dados, foi feita uma validação a partir do conjunto de dados de teste, previamente separados. Para ambos os testes, os modelos obtiveram resultados similares. Por fim, a partir dos coeficientes obtidos a partir dos estimadores de máxima verossimilhança dos modelos, foi possível obter o mapa de distribuição potencial para cada um deles. Esses foram comparados com os obtidos por meio do Modelo de Máxima Entropia.

Como o mapa gerado pelo modelo com distribuições de probabilidade *a priori* informativas apresentou maior semelhança, concluiu-se que o modelo de regressão logística Bayesiana é viável para a tarefa de modelagem de distribuição de espécies. Além disso, a incorporação de conhecimentos prévios no modelo mostra-se de grande relevância para

a obtenção de resultados próximos aos de modelos clássicos. Para este trabalho, as distribuições de probabilidade *a priori* foram obtidas por meio de estudos prévios, porém o uso de conhecimentos de especialistas mostrou-se de grande potencial na literatura e pode ser avaliado futuramente. Ademais, também é sugerida a avaliação do emprego do modelo para outras espécies, em especial as consideradas raras ou com risco de extinção.

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.

Di Lorenzo, B., Farcomeni, A., and Golini, N. (2011). A bayesian model for presence-only semicontinuous data, with application to prediction of abundance of taxus baccata in two italian regions. *Journal of Agriculture Biological and Environmental Statistics*, 16:339–356.

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters*, 7:509–520.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511.

Gelman, A. and Rubin, D. B. (1995). Avoiding model selection in bayesian social research. *Sociological Methodology*, 25:165–174.

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

Hegel, T. M., Cushman, A., Evans, J., and Huetmann, F. (2010). *Spatial Complexity, Informatics and Wildlife Conservation*, chapter Current State of the Art for Statistical Modelling of Species Distributions. Springer.

Hutchinson, G. E. (1991). Population studies: Animal ecology and demography. *Bulletin of Mathematical Biology*, 53(1-2):193–213.

ICMBio (2021). Portal da biodiversidade do instituto chico mendes de conservação da biodiversidade. https://portaldabiodiversidade.icmbio.gov.br/portal/. Acesso em: 2021-04-21.

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

Kunhert, P. M., Martin, T. G., and P, G. S. (2010). A guide to eliciting and using expert knowledge in bayesian ecological models. *Ecology Letters*, 13:900–914.

Low-Choy, S. L., O'Leary, R., and Mergensen, K. (2009). Elicitation by design in ecology: using expert opinion to inform priors for bayesian statistical models. *Ecology*, 90(1):265–277.

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, T. G., Kuhnert, P. M., Mengersen, K., and Possingham, H. P. (2005). The power of expert opinion in ecological models using bayesian methods: Impact of grazing on birds. *Ecological Applications*, 15:266–280.

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.

Niamir, A. (2019). Incorporating knowledge uncertainty into species distribution modelling. *Biodiversity and Conservation*, 28:571–588.

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

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.

Salvatier, J., Wieck, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *Peer J Computer Science*.