Previsão de ocupação de anuros usando sensoriamento de variáveis ambientais e modelos Autocodificadores

Nabson P. S. da Silva, Juan G. Colonna, Marco Cristo

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Av. Rodrigo Octávio, 6200, CEP: 69067-005, Manaus, AM, Brasil

{nabson.paiva, juancolonna, marco.cristo}@icomp.ufam.edu.br

Resumo. *Neste trabalho apresentamos um modelo de previsão da distribuição geográfica de anuros baseado em uma Rede Neural Autocodificadora. O problema de previsão é tratado como uma tarefa de classificação One-Class, em que a localização conhecida dos indivíduos é representada através de suas coordenadas espaciais, temporais e variáveis meteorológicas correlacionadas. O modelo mostrou alto grau de acuidade quando utilizado para prever a ocorrência de sapos da espécie Bufo Americanus no sul do Canadá. O método proposto é computacionalmente leve e pode ser acoplado a redes de sensores para monitoramento ambiental.*

Abstract. *We present a model for predicting geographic distribution of anurans based on Autoencoder. The prediction problem is treated as an One-Class classification task, in which each sample is represented by its geographic coordinates, temporal and correlated meteorological variables. Our model shows a high degree of accuracy when used to predict the occurrence of the Bufo Americanus species in southern Canada. The proposed method is computationally inexpensive and can be coupled to sensor networks for environmental monitoring.*

1. Introdução

A distribuição geográfica de uma espécie permite observar aspectos relevantes associados à sua conservação como suas preferências habitacionais, o quão ameaçada de extinção ela está e as condições bióticas e abióticas que ajudam na sua sobrevivência [Phillips et al. 2004]. Além disso, estudos recentes mostram que a redistribuição de espécies pode ser utilizada para determinar a possibilidade de se deixar a própria natureza trabalhar em sua conservação (*rewilding*) sob os futuros cenários de mudança climática [Jarvie and Svenning 2018].

Modelos de distribuição geográfica de espécies podem ainda auxiliar estudos biológicos ao facilitar processos de monitoração e tornar mais efetivos trabalhos de campo, reduzindo assim o tempo de estudo e demanda por recursos econômicos. Tais vantagens têm estimulado o desenvolvimento de modelos de previsão da ocorrência geográfica de espécies que explorem os seus aspectos determinantes como: a intervenção humana, as estações do ano, a disponibilidade de recursos, os fatores abióticos e bióticos. Muitos destes podem ser capturados, direta e indiretamente, por meio de dados espaciais e meteorológicos obtidos por sensores ambientais ao longo do tempo.

Em nossa pesquisa tratamos o padrão de distribuição de espécies como um problema de detecção de anomalias ou de classificação *One-Class* [Khan and Madden 2014].

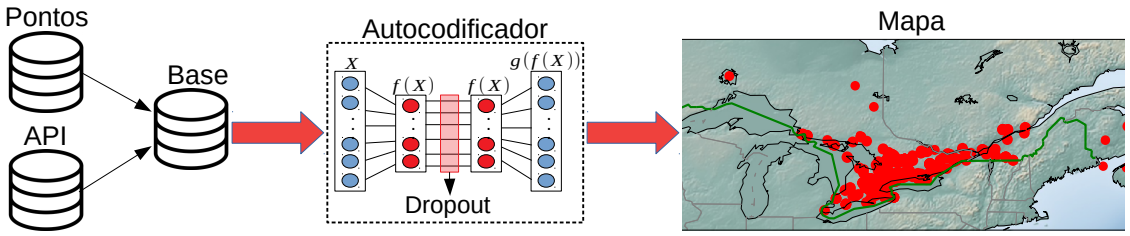


Figura 1. Método proposto para previsão de ocupação de espécies. Os pontos vermelhos identificam a distribuição geográfica da espécie American Toad. As bases de dados "Pontos" e "API" representam as observações humanas e as consultas aos sensores meteorológicos, respectivamente.

Neste problema, o objetivo é prever se as instâncias pertencem ou não à classe "alvo", ou seja, se são ou não casos anômalos. Esse tipo de classificação é comum quando não existem amostras da classe negativa, como é o caso dos estudos ecológicos de distribuição de espécies, nos quais não existem uma base de dados com observações indicando pontos onde o animal não foi avistado.

Propomos o uso do método de classificação *One-Class* para estimar a probabilidade de ocorrência de indivíduos de uma espécie de anuros por meio do uso de autocodificadores (*Autoencoder* - AE), os quais são redes neurais usadas como detectores de anomalias. Essas redes são treinadas para reconstruir a instância de entrada na saída [Géron 2017] sendo, portanto, um método não supervisionado (ele não depende de rotulação manual). Considerando que o AE tenta reconstruir cada entrada com o mínimo erro possível, uma amostra que apresenta alto erro de reconstrução deve representar uma instância muito distinta de todas as observadas no treino e, portanto, uma anomalia [Mazhelis 2006].

No contexto de distribuição de espécies, como aborda este trabalho, cada instância representa a geolocalização espacial de um individuo em determinado momento no tempo combinada a dados meteorológicos capturados por sensores ambientais associados àquela localização. Dada uma instância, o erro de reconstrução do AE pode ser usado para estimar a probabilidade de observação do individuo de forma a gerar um mapa que indica onde há maiores chances da espécie ser avistada, como ilustrado na Figura 1.

Aplicamos nossa abordagem em uma espécie de sapos da espécie American Toad (*Bufo americanus*) [AmphibiaWeb 2019] de maneira a gerar um mapa de distribuição potencial com base na probabilidade de ocorrência dos espécimes no Canadá. Essa espécie, endêmica da América do Norte, é conhecida por ser numerosa e encontrada em diferentes tipos de habitat, não apenas em torno de rios e lagos [Society 2018]. Obtivemos vinte e uma variáveis ambientais por meio de estações meteorológicas distribuídas no Canadá, que foram pré-processadas, normalizadas e selecionadas para posteriormente serem utilizadas como instâncias de entrada do AE. O modelo mostrou grande acuidade quando comparamos suas previsões com dados reais de distribuição dos anuros.

Este artigo encontra-se organizado como segue. A Seção 2 detalha a importância do estudo da distribuição de espécies, os problemas de classificação de uma classe e a rede neural utilizada. A Seção 3 descreve o método proposto para a solução do problema. A Seção 4 apresenta a coleta da base de dados, o pré-processamento das variáveis, o trei-

namento do modelo proposto e os resultados obtidos. Por fim, na Seção 5 apresentamos nossas conclusões e novos caminhos para a pesquisa.

2. Fundamentos teóricos

A criação de modelos preditores de distribuições geográficas de espécies com base em dados ambientais correlacionados de sítios de ocorrência conhecida constitui uma importante técnica em biologia analítica. Ela tem aplicações em conservação e planejamento de reservas, ecologia, evolução, manejo de espécies invasoras e outros campos [Phillips et al. 2006], o que motivou a pesquisa na criação de modelos de inferência.

O habitat escolhido por uma espécie depende tanto do clima quanto da disponibilidade de recursos naturais, relação com as outras espécies presentes, além de outros fatores. Contudo, o uso de dados meteorológicos correlacionados aliados a variabilidade temporal mostrou ser capaz de de definir a adequação do habitat [Bateman et al. 2012]. Com base nisso, esta pesquisa utiliza os dados de uma estação meteorológica e de um *website* colaborativo que registra as coordenadas e a data de aparecimento de animais para inferir quando e onde há mais chances de eles serem vistos novamente.

Em problemas de classificação, o objetivo do modelo é decidir a qual classe predefinida o dado pertence. Para isso é necessário ter várias amostras para cada uma das classes. Entretanto, em problemas *One-Class*, existe apenas uma classe conhecida caracterizada pelos dados na base enquanto a classe negativa (a qual pertencem as instâncias fora da classe alvo) possui poucos ou nenhum exemplo, o que é insuficiente para treinar um modelo multi-classe [Khan and Madden 2014]. Como as bases de estudo de distribuição de espécies possuem apenas ocorrências dela, tratamos o problema como *One-Class*.

Por serem treinadas unicamente com amostras positivas, As Rede Neurais Autocodificadoras podem ser usadas para tarefas *One-Class* de forma a aprender uma representação distribuída do conjunto das amostras da classe alvo [Mazhelis 2006]. Redes desse tipo são capazes de aprender uma representação eficiente dos dados de entrada sem supervisão. Para isso, as camadas escondidas diminuem a dimensão do vetor de entrada para uma versão mais representativa dele, processo chamada codificação. A partir dessa nova representação comprimida da informação original, a rede tenta reconstruir o vetor características originais, etapa chamada decodificação [Géron 2017].

3. Método proposto

Como observado, podemos tratar o problema de previsão de ocorrência como um problema de detecção de anomalia. Para tanto, nós usamos uma rede autocodificadora para prever a probabilidade de aparecimento dos espécimes em novas coordenadas sob determinadas condições meteorológicas.

Ao treinar um AE, ele aprenderá a reconhecer padrões entre as amostras de forma a conseguir codificá-los e decodificá-los. Dado $\mathbf{x}$ um vetor de variáveis ambientais, coordenadas geográficas e data, a função de codificação é $f(x) = \sigma(W_i\mathbf{x} + b_i)$, em que W_i são os pesos das camadas codificadoras do AE encarregados de diminuir a dimensão do dado original, b_i é vetor de viés e σ é a função de ativação da camada. De forma análoga, a função de decodificação é $g(f(\mathbf{x})) = \sigma(W_o f(\mathbf{x}) + b_o)$, onde W_o são os pesos da camada de reconstrução e b_o os viés.

Em termos de dimensões matriciais, a matriz W_i possui as mesmas dimensões que a matriz W_o^T transposta, permitindo projetar o vetor original $\mathbf{x}$ em um espaço vetorial reduzido e, posteriormente, aplicar $g(\cdot)$ para projetá-lo novamente no espaço vetorial original. Na codificação e na decodificação existe a função de ativação σ, cujo objetivo principal é permitir que as projeções vetoriais sejam não lineares.

Finalmente, o AE é treinado ajustando as matrizes W_i e W_o, de forma que o erro de reconstrução $L = ||\mathbf{x} - g(f(\mathbf{x}))||_2$ seja minimizado. Note-se que $|| \cdot ||$ é a norma euclidiana entre os vetores. Assim, vetores novos reconstruídos com alto erro L devem representar instâncias muito distintas das observadas no treinamento e, portanto, essencialmente diferentes daquelas na classe positiva aprendida pelo AE.

Ao quantificar o erro de reconstrução de uma instância $\mathbf{x}$, é possível estimar a probabilidade de uma instância similar a ela ter sido observada durante o treinamento do modelo. Para tanto, nós propomos utilizar um Kernel de Densidade Gaussiana [Fukunaga 2013] para mapear erros de reconstrução $(x_i - \hat{x}_i)^2$ para valores de probabilidades ($0 \leq P(y = 1|\mathbf{x}) \leq 1$). Desta forma, dado o erro de reconstrução, é possível estimar as chances de encontrar indivíduos da espécie alvo em uma localidade específica, sob determinadas circunstâncias climáticas. Em particular, a probabilidade condicional de presença da espécie $P(y = 1|\mathbf{x})$ é obtida aplicando a função de Kernel Gaussiano em cada característica e tomando a média aritmética delas de acordo com:

$$P(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} e^{-(x_i - \hat{x}_i)^2 / 2\sigma}, \tag{1}$$

onde σ corresponde à variância da estimativa e n ao número de caraterísticas ambientais. Nos experimentos que realizamos neste estudo, usamos $\sigma = 0,01$ e $n = 21$.

4. Metodologia experimental

Para avaliar nosso modelo, construímos uma base de dados utilizando uma API remota que fornece medições de sensores ambientais. A base foi pré-processada, o que incluiu o processo de seleção de atributos, e utilizada para o treinamento do AE de forma a gerar um mapa de probabilidades de distribuição espacial das espécies. Tais etapas são detalhadas nas subseções a seguir.

4.1. Coleta da base

A base de dados foi minerada do *website* colaborativo [FrogWatch 2018] e contém a data, latitude e longitude de onde os anuros foram avistados. Essa plataforma serve como espaço para que voluntários ao redor do mundo possam registrar onde e quando avistaram diferentes espécies animais e diversos tipos de plantas, vermes, gelo e outros. Dentre todas as espécies registradas no FrogWatch, a espécie American Toad foi escolhida para nosso estudo por possuir o maior numero de amostras, sendo no total 3686 registros. O lado esquerdo da Figura 1 apresenta os pontos que compõem a base de dados, onde os espécimes foram avistados.

Após a coleta das obervações temporais geo-localizadas do FrogWatch, a API [WeatherUnderground 2018] foi consultada para obter os registros históricos climáticos dos sensores meteorológicos distribuídos no Canadá. Esta API permite consultar dados climáticos históricos selecionando uma coordenada geográfica e uma data

específica. Assim, conseguimos atribuir a cada observação da base de amostras um vetor de caraterísticas ambientais. A lista destas variáveis é descrita na próxima seção.

4.2. Pré-processamento do dados

O tratamento da base iniciou-se pela limpeza de características com poucos dados, deixando apenas as que possuem mais de 50% e que estão representadas no sistema métrico brasileiro. Para valorizar a estação do ano em que os sapos aparecem com maior frequência, a variável "data" foi convertida em categórica utilizando unicamente os meses do ano. As características "latitude" e "longitude" foram convertidas de graus para radianos para facilitar a representação dos pontos no sistema de coordenadas polares.

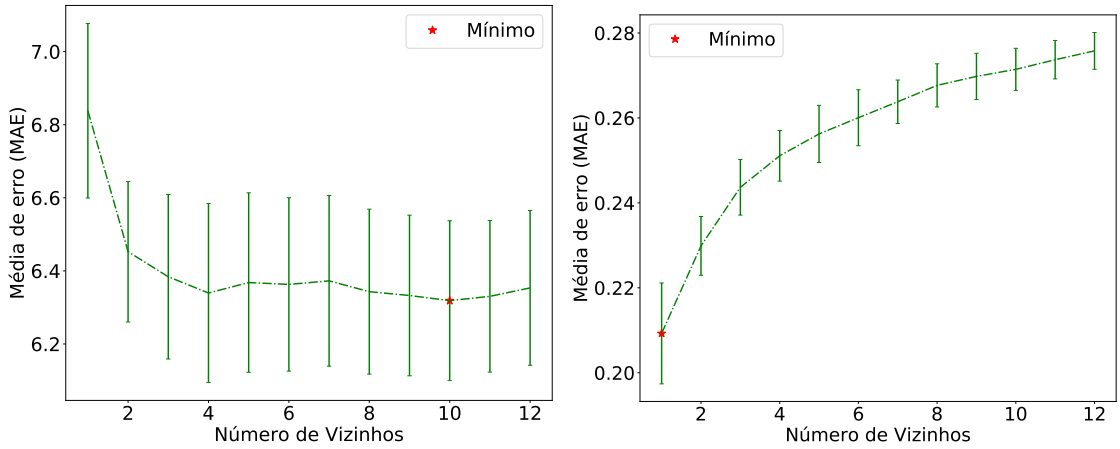
O preenchimento das variáveis com valores faltantes (*Not-a-Numbers* - NaN) não pode ser realizado usando métodos tradicionais, como o preenchimento com a média da variável. Isto deve-se a que os dados meteorológicos dependem da data da medição e da localização onde foram medidos. Portanto, para preencher os NaN, optamos por usar o método de regressão K-vizinhos Mais Próximos (*k*NN) [Chen et al. 2018]. Com esta regressão, os valores faltantes para cada variável foram imputados usando pontos geograficamente próximos em datas similares.

Tendo em vista que para cada variável será necessário utilizar um modelo de regressão *k*NN diferente, tivemos que realizar uma validação cruzada para calcular o erro médio e o intervalo de confiança para cada possível valor de k (número de vizinhos), variando entre $1 \leq k \leq 12$. A escolha do melhor k deu-se ao analisar o menor erro absoluto médio (*Mean Absolute Error* - MAE) [Géron 2017], bem como os intervalos de confiança sem sobreposição para verificar a significância estatística do valor k que minimiza o erro.

A Figura 2 apresenta dois gráficos que ilustram a variação do MAE e do intervalo de confiança em relação ao parâmetro k. O gráfico 2(a) mostra a análise da característica "maxVelocidadeVento", na qual a maioria dos intervalos de confiança possuem interseção. Neste caso, como qualquer valor entre $3 \leq k \leq 12$ seria satisfatório, $k = 10$ foi escolhido. Já no gráfico 2(b), que apresenta a característica "chuva", observamos que o erro mínimo não possui qualquer interseção com os outros intervalos de confiança. Assim, para este caso, optamos por $k = 1$. Essa análise foi repetida para todas as características da base de forma a obtermos valores adequados de k para cada caso.

A Tabela 1 resume os erros obtidos para cada característica pelo procedimento descrito anteriormente. Os valores destacados em negrito apresentaram empate estatístico. As características restantes após o pré-processamento foram renomeadas conforme: "data": f0; "latitude": f1; "longitude": f2; "trovao": f3; "neve": f4; "minHumidade": f5; "grausResfriamento": f6; "maxPontoOrvalho": f7; "grausAquecimento": f8; "mediaPressao": f9; "chuva": f10; "minVelocidadeVento": f11; "nevoa": f12; "humidade": f13; "graus": f14; "minPontoOrvalho": f15; "minPressao": f16; "maxTemperatura": f17; "mediaVelocidadeVento": f18; "maxPressao": f19; "granizo": f20; "mediaPontoOrvalho": f21; "maxVelocidadeVento": f22; "precipitacao": f23, "mediaTemperatura": f24; "mediaDirecaoVento": f25; "maxHumidade": f26; e "minTemperatura": f27.

Com a base pré-processada, é preciso que todos os dados estejam na mesma escala para que a rede neural consiga convergir. Para isso, foi aplicado o método *Robust Scaler* seguido pelo *MinMax Scaler* [Theodoridis and Koutroumbas 2008]. O primeiro



(a) Caraterística "maxVelocidadeVento". (b) Caraterística "chuva".

Figura 2. Análise do erro médio e do intervalo de confiança em relação ao número de vizinhos k.

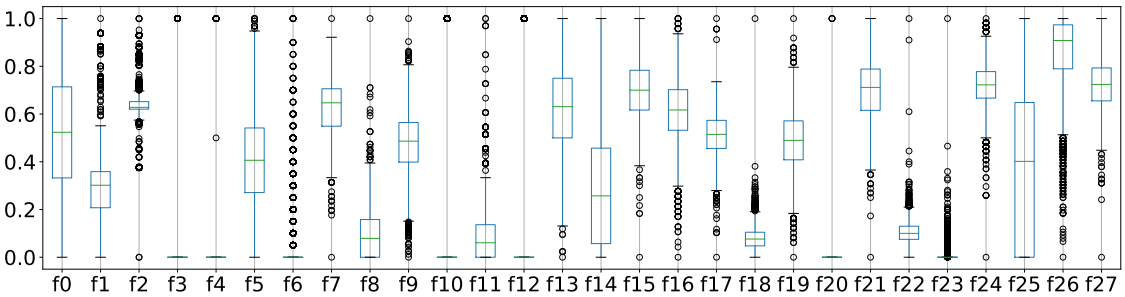


Figura 3. *Boxplot* das características normalizadas resultantes do pré-processamento.

é utilizado para prevenir que *outliers* interfiram na distribuição da variável, enquanto que o segundo é aplicado para normalizar a escala das caraterísticas para valores entre $0 \leq x_i \leq 1$.

4.3. Seleção de caraterísticas para o AE

A Figura 3 mostra o *Boxplot* das características após a limpeza dos dados descrita na seção anterior. Neste gráfico, percebe-se a distribuição dos valores de cada caraterística. Algumas destas estão concentradas ao redor de um único valor, como o caso de "f3", na qual a maioria dos valores são iguais a 1. Variáveis como este tipo de distribuição prejudicam o aprendizado, pelo fato de não haver variação nos padrões a serem aprendidos.

Para detectar variáveis pouco informativas, aplicamos o critério interquartil $IR = Q_3 - Q_1$ [Lorena et al. 2015], onde Q_1 e Q_3 são os primeiro e terceiro quartis, respectivamente. Este critério analisa a diferença entre os extremos de Q_1 e Q_3 para identificar variáveis cuja distribuição não possui uma variança significativa, como é o caso de f3, f4, f6, f9, f13, f21 e f24, que obtiveram $IR \approx 0$, resultado observado na Figura 3. Consequentemente, estas caraterísticas foram eliminadas da base para evitar o uso de variáveis ambientais que não contribuem com informações úteis e confundem o processo de aprendizado de nossa rede neural. Com isso, as características restantes que foram utiliza-

Tabela 1. Escolha do melhor valor do parâmetro k para cada caraterística.

Carac. \ k	1	2	3	4	5	6	7	8	9	10	11	12
f3	**0.05**	-0.06	-0.07	-0.07	-0.07	-0.07	-0.07	-0.08	-0.08	-0.08	-0.08	-0.08
f4	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
f5	**-12.0**	**-11.8**	**-11.9**	**-12.0**	-12.1	-12.2	-12.4	-12.4	-12.6	-12.6	-12.7	-12.8
f6	**-1.04**	**-1.10**	-1.14	-1.18	-1.22	-1.25	-1.28	-1.31	-1.34	-1.36	-1.39	-1.14
f7	**-2.69**	**-2.69**	**-2.72**	**-2.77**	-2.83	-2.88	-2.92	-2.96	-3.00	-3.03	-3.07	-3.10
f8	**-3.34**	**-3.44**	-3.54	-3.63	-3.72	-3.80	-3.89	-3.96	-4.03	-4.08	-4.13	-4.16
f9	**-3.39**	**-3.39**	**-3.47**	**-3.59**	-3.70	-3.78	-3.85	-3.91	-3.95	-3.99	-4.03	-4.08
f10	**-0.20**	-0.22	-0.24	-0.25	-0.25	-0.26	-0.26	-0.26	-0.26	-0.27	-0.27	-0.27
f11	**-2.59**	**-2.57**	**-2.58**	**-2.60**	**-2.62**	-2.64	-2.66	-2.67	-2.68	-2.69	-2.69	-2.70
f12	**-0.07**	-0.07	-0.07	-0.08	-0.08	-0.08	-0.08	-0.08	-0.09	-0.09	-0.09	-0.09
f13	-9.34	**-8.92**	**-8.98**	**-9.090**	-9.18	-9.33	-9.50	-9.61	-9.72	-9.81	-9.89	-9.96
f14	**-3.44**	**-3.55**	-3.65	-3.75	-3.85	-3.93	-4.02	-4.11	-4.19	-4.26	-4.32	-4.38
f15	**-2.90**	**-2.85**	**-2.91**	-2.98	-3.04	-3.11	-3.17	-3.21	-3.25	-3.29	-3.32	-3.36
f16	**-3.62**	**-3.64**	**-3.71**	-3.84	-3.95	-4.03	-4.10	-4.16	-4.21	-4.25	-4.30	-4.35
f17	**-3.02**	**-3.01**	-3.09	-3.17	-3.23	-3.31	-3.38	-3.44	-3.50	-3.54	-3.57	-3.60
f18	-3.47	**-3.31**	**-3.26**	**-3.27**	**-3.28**	**-3.29**	**-3.30**	**-3.30**	**-3.31**	**-3.31**	**-3.31**	**-3.32**
f19	**-3.12**	**-3.15**	**-3.23**	-3.34	-3.41	-3.49	-3.56	-3.62	-3.66	-3.69	-3.74	-3.78
f20	**-0.00**	**-0.00**	**-0.00**	**-0.00**	**-0.00**	**-0.00**	**-0.00**	**-0.00**	**-0.00**	**-0.00**	**-0.00**	**-0.00**
f21	**-2.56**	**-2.55**	**-2.60**	-2.67	-2.73	-2.80	-2.85	-2.90	-2.95	-2.98	-3.01	-3.05
f22	-6.83	**-6.45**	**-6.38**	**-6.33**	**-6.36**	**-6.36**	**-6.37**	**-6.34**	**-6.33**	**-6.31**	**-6.33**	**-6.35**
f23	**-1.45**	**-1.49**	**-1.45**	**-1.46**	**-1.45**	**-1.46**	**-1.45**	**-1.47**	**-1.47**	**-1.48**	**-1.48**	**-1.49**
f24	**-2.44**	**-2.49**	-2.54	-2.60	-2.65	-2.69	-2.75	-2.80	-2.85	-2.88	-2.92	-2.95
f25	**-77.3**	**-78.6**	-79.8	-80.4	-80.9	-81.4	-82.1	-82.5	-82.9	-83.1	-83.3	-83.5
f26	**-7.31**	**-7.13**	**-7.13**	**-7.19**	**-7.27**	-7.34	-7.37	-7.40	-7.45	-7.48	-7.52	-7.56
f27	**-3.00**	**-2.95**	**-2.97**	**-3.01**	**-3.06**	-3.10	-3.16	-3.21	-3.25	-3.27	-3.29	-3.32

das para treinar a rede referem-se a temperatura, ponto de orvalho, velocidade do vento, pressão e grau de aquecimento.

4.4. Treinamento da Rede Autocodificadora

Duas estratégias de separação da base em treino e teste foram aplicadas. Na primeira, chamada B_1, uma amostra aleatória de 20% dos dados (teste 1) é usada como conjunto de teste para posteriormente avaliar a convergência da rede simulando novas observações. Na segunda, chamada B_2, as amostras datadas do mês de maio (teste 2), época em que os sapos aparecem com maior frequência, são usadas como teste, de forma a verificar o desempenho do modelo durante o mês de maior atividade da espécie.

Foram comparadas cinco arquiteturas de redes autocodificadoras, denominadas AE_1, AE_2, AE_3, AE_4 e AE_5 (cf. Figura 4). A Tabela 2 mostra o erro de treinamento e validação nas bases B_1 e B_2. O erro quadrático médio (MSE) foi escolhido para monitorar a convergência das redes durante o treinamento [Géron 2017]. Quanto menor é o MSE, melhor é o modelo. Observa-se que em ambas bases de dados a rede que apresentou menor MSE no conjunto de validação foi a AE_5, portanto esta foi escolhida como modelo final.

O modelo final AE_5 possui duas camadas escondidas, como mostra a Figura 4. A codificação da primeira camada oculta diminui a dimensão do vetor de entrada pela metade e realiza a transformação não linear usando a função de ativação tangente hiperbólica (Tanh). Após a primeira camada oculta, outra camada semelhante foi concatenada com o objetivo de aumentar a representatividade dos padrões no espaço latente. Uma camada com 10% de Dropout foi adicionado entre as duas camadas ocultas para evitar um possível sobre-ajuste (*overfiting*). Por último, a camada decodificadora final recria o vetor de en-

Tabela 2. Comparação entre Arquiteturas de Rede Autocodificadoras. Os melhores valores foram destacados em negrito.

Base	AE_1	AE_2	AE_3	AE_4	AE_5
B_1: Treino	0.01056	0.00995	0.00365	0.00212	**0.00191**
B_1: Validação	0.00649	0.00651	0.00232	0.00094	**0.00090**
B_2: Treino	0.00952	0.00970	0.00297	0.00172	**0.00158**
B_2: Validação	0.00626	0.00786	0.00199	0.00108	**0.00106**

trada utilizando a função de ativação Sigmoid [Géron 2017].

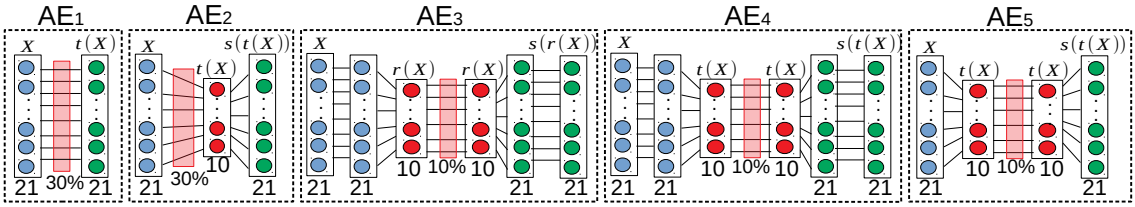


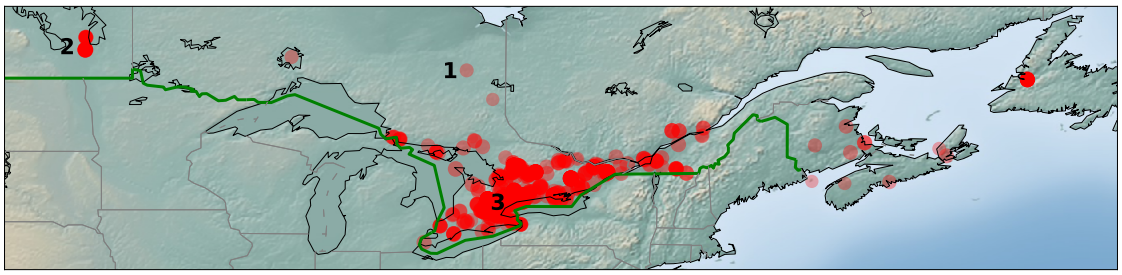
Figura 4. As camadas azuis representam as entradas, as vermelhas o espaço latente e as verdes as saídas decodificadas. Os números indicam a quantidade neurônios. Blocos vermelhos opaco representam o Dropout com sua porcentagem de regularização. As funções de ativação são: *Relu* $r(X)$, *Tanh* $t(X)$ e *Sigmoid* $s(X)$. Camadas sem ativação representam o *Batch normalization*.

AE_1 usa a função *Tanh* e Dropout de 30% entre a entrada e a saída. AE_2 codifica a a entrada para 10 neurônios com Dropout de 30% e função *Tanh*, e decodifica o vetor latente com a função *Sigmoid*. AE_3 e AE_4 diferenciam-se pelas funções de ativação utilizadas: *Relu* e *Tanh*, respectivamente. Ambas possuem uma camada de Normalização em Batch [Géron 2017] após a entrada e antes da saída e codificam a entrada para 10 neurônios. Após codificar a entrada, outra camada com mesmo número de neurônios e função de ativação é colocada com Dropout de 10% entre elas.

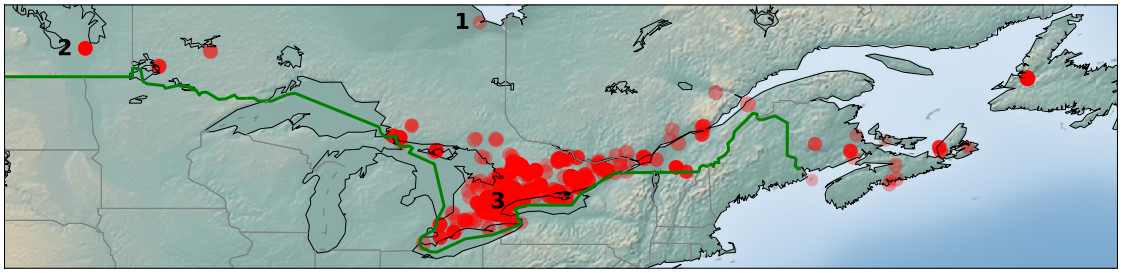
4.5. Resultados

Nesta seção foram utilizadas as bases de teste B_1 e B_2 mencionadas na seção anterior. Para cada amostra destas bases a probabilidade de ocupação da espécie foi inferida aplicando a Equação 1. Com isto, foi possível gerar os mapas de distribuição apresentados nas Figuras 5(a) e 5(b). Nesses mapas, o diâmetro e a transparência de cada ponto é proporcional a $P(y = 1/x)$. Assim, os pontos coordenados com maior probabilidade de aparecimento de sapos são maiores e menos transparentes. Os mapas correspondem às latitudes máxima e mínima de 52º e 41º, respectivamente, e longitudes -54º e -100,5º.

Em ambas as Figuras, 5(a) e 5(b), podemos observar que o ponto 1 possui um diâmetro menor que o ponto 2 e também é mais opaco, o que simboliza as chances maiores que existem de encontrar a espécie no ponto 2 em relação ao ponto 1. O ponto 3 representa a área com maior chance de aparecerem sapos da espécie American Toad, ao sul de Ontário. Os resultados comparativos entre a amostragem aleatória e a amostragem realizada unicamente no mês de maio mostram que esta espécie é predominante na temporada de primavera e no inicio do verão. Estas observações confirmam a capacidade correta de prever a ocorrência de sapos American Toad usando o modelo proposto.



(a) Pontos de observação aleatórios.



(b) Pontos de observação no mês de maio.

Figura 5. Mapas de distribuição das espécies usando: (a) B_1 e (b) B_2.

5. Conclusão

Esta pesquisa apresentou um modelo para estimar a distribuição espacial de uma espécie de anuros através de dados de sensores meteorológicos. O método proposto utiliza redes neurais artificias para aprender sobre as amostras da base em que os sapos foram avistados. Com o modelo proposto foi possível gerar um mapa de distribuição espacial de espécimes American Toad.

Tendo em vista que a base possui apenas pontos positivos, utilizamos uma Rede Autocodificadora para resolver o problema de classificação *One-Class* com base no erro de reconstrução. Por meio disso, foi possível inferir o quão discrepantes são as novas observações em relação aos dados de treinamento, o que indica se estas fazem parte ou não da classe alvo.

O método tem complexidade assintótica $O(2 * nmp)$, em que m é o número de características no vetor de entrada (21), n é igual a m e p é a dimensão da camada oculta (10). A inferência para um vetor de entrada já tratado faz cerca de 520 multiplicações, o que tem um baixo custo computacional e justifica o uso da rede em sensores de monitoramento. Sendo assim, a abordagem proposta pode ser implementada em sensores em campo capazes de predizer onde e quando há mais chances de uma espécie aparecer sem precisar de um servidor externo.

Como trabalho futuro, nós pretendemos utilizar Redes Autocodificadoras Variacionais que, por aprender distribuições de probabilidades das variáveis latentes, possibilitariam gerar novas amostras em regiões de alta probabilidade de aparecimento dos espécimes [Géron 2017]. Também pretendemos expandir a pesquisa para novas espécies de animais, analisar a distribuição de acordo com as estações do ano e incorporar métodos de seleção de características mais robustos.

Referências

AmphibiaWeb (2019). American toad. Disponível em: https://amphibiaweb.org. Acesso em: 11 de março de 2019.

Bateman, B. L., VanDerWal, J., and Johnson, C. N. (2012). Nice weather for bettongs: using weather events, not climate means, in species distribution models. *Ecography*, 35(4):306–314.

Chen, G. H., Shah, D., et al. (2018). Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends in Machine Learning*, 10(5-6):337–588.

FrogWatch (2018). Frogwatch: Enganging citizens in science. Disponível em: https://www.naturewatch.ca/frogwatch/. Acesso em: 13 de dezembro de 2018.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Elsevier.

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media, 1st edition.

Jarvie, S. and Svenning, J. C. (2018). Using species distribution modelling to determine opportunities for trophic rewilding under future scenarios of climate change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1761).

Khan, S. S. and Madden, M. G. (2014). One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374.

Lorena, L. H. N., Carvalho, A. C. P. L. F., and Lorena, A. C. (2015). Filter feature selection for one-class classification. *Journal of Intelligent & Robotic Systems*, 80(1):227–243.

Mazhelis, O. (2006). One-class classifiers: a review and analysis of suitability in the context of mobile-masquerader detection. *South African Computer Journal*, 2006(36):29–48.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259.

Phillips, S. J., Dudík, M., and Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning*, page 83. ACM.

Society, C. H. (2018). American toad. Disponível em: http://canadianherpetology.ca/species/species_page.html?cname=American%20Toad. Acesso em: 13 de março de 2019.

Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition, Fourth Edition*. Academic Press, Inc., 4th edition.

WeatherUnderground (2018). Api: Weather underground. Disponível em: https://www.wunderground.com/weather/api. Acesso em: 15 de dezembro de 2018.