Aluno: RAFAEL ROCHA MATTAZIO
Professores: Hugo Fernando Maia Milan e Alex Sandro Campos Maia
Disciplina: TEAMAZ2 – Aprendizado de Máquina Aplicada à Zootecnia II

Relatório

Previsão de áreas de apreensão de aves ameaçadas usando aprendizado de máquina

26 de fevereiro à 16 de março de 2022
São Carlos – SP

1 Objetivo

Dentro da motivação de coibir a captura de aves silvestres, sob a hipótese de haver um padrão entre espécie apreendida, local e data, o objetivo é realizar forecasting de locais de apreensão de aves ameaçadas para orientar ações de fiscalização.

2 Dataset

São usados dados dos conjuntos:

1. base de dados abertos de apreensões do IBAMA (termo de apreensão e espécime apreendido)
2. lista de espécies ameaçadas e quase ameaçadas divulgas pelo ICMBio.
3. lista das aves do Brasil do Comitê Brasileiro de Registros Ornitológicos (CBRO)

Dessas bases foram selecionados os seguintes dados:

- `SEQ_TAD`: Chave que identifica o termo de fiscalização, a qual permite correlacionar as bases de dados diferentes do Ibama;
- `DAT_TAD`: Data em que o termo foi lavrado (ISO 8601);
- `SIG_UF`: Sigla da unidade da federação referente ao local da apreensão.
- `NUM_LONGITUDE_TAD`: Longitude do local da apreensão;
- `NUM_LATITUDE_TAD`: Latitude do local da apreensão;
- `NOME_CIENTIFICO/Nome.do.táxon..sem.autoria./Espécie`: Nome científico do espécime apreendido;
- `Gênero`: classificação taxonômica;
- `Ordem`: classificação taxonômica;
- `Família`: classificação taxonômica;
- `Categoria.validada`: Estado de conservação da ave;
- `QTD_ESPECIME_APREENDIDO`: Quantidade de espécimes apreendidos

2.1 Tratamento dos dados

Antes da utilização dos dados nos algoritmos, foi feito o seguinte pré-processamento, partindo de um total inicial de 47331 linhas de dados referentes à apreensões de aves:

1. remoção de `NA`: 17427 linhas restantes;
2. segmentação da data em ano e mês;
3. arredondamento dos valores decimais de latitude e longitude conforme estratégia de modelagem:
    - 0 casas decimais = área de aprox. 100km x 100km;
    - 1 casa dec. = cidade média;
    - 2 c.d. = bairro;
    - 3 c.d. = quarteirão.

4. filtragem de outliers e intervalos pouco representados, resultando em 14712 linhas restantes:
    - 2013 < ano < 2019;
    - 0 < quantidade apreendida < 200;
    - -73.992 < longitude < -34.765;
    - -33.753 < latitude < 5.280:
5. remoção dos casos cujo nome científico apresentado no termo de apreensão não constava lista do CBRO ou do ICMBio: 11078 linhas restantes;
6. agrupamento das linhas segundo casos únicos de covariáveis para cálculo da soma da quantidade apreendida de cada espécie: as linhas restantes dependem das variáveis de entrada adotadas em cada estratégia de modelagem;
7. consideração do estado de conservação mais crítico dentro do nível taxonômico.
8. divisão do dataset em treino e teste utilizando como segmentação o ano de 2018, uma vez que se trata de uma série temporal, que resultou numa parcela de cerca de 20% das linhas para teste[1].

3 Modelagem

Esses dados são utilizados da seguinte maneira:

- *Input*: mês, ano, lat/long ou estado, ordem ou família ou gênero, estado de conservação;
- *Output*: quantidade apreendida;

com estratégias definidas por diferentes complexidades:

1. simplificada: dados agregados por estado (sem distinção de coordenadas), ano de ocorrência e categoria de conservação, com inclusão de casos com zero apreensões para completude da série temporal, ou seja: `DAT_TAD_year + SIG_UF + Categoria.validada = QTD_ESPECIME_APREENDIDO`;
2. intermediária: dados agregados por par de coordenadas com arredondamento para uma casa decimal (compreendendo área aproximada de uma cidade média), ano e mês de ocorrência, hierarquia taxonômica até gênero e categoria de conservação, porém sem inclusão de zeros, portanto: `DAT_TAD_year + DAT_TAD_month + NUM_LONGITUDE_TAD + NUM_LATITUDE_TAD + Gênero + Categoria.validada = QTD_ESPECIME_APREENDIDO`;
3. refinada: coordenadas somente com algarismo inteiros (compreendendo área aproximada 100km x 100km), ano e mês de ocorrência, hierarquia taxonômica até família e categoria de conservação, com inclusão de casos com zero apreensão para completude da série temporal.

[1] Cross-validation não foi implementado, mas seguiria uma lógica sequencial conforme sugerido por Hyndman e Athanasopoulos.

As técnicas de aprendizado empregadas foram:

- modelo de valor médio estratificado por estado ou coordenadas geográficas;
- modelo linear generalizado (glm);
- florestas aleatórias (rf);
- Gradient boosting machines (gbm);
- redes neurais artificiais (ann) ou deep learning (dl).

Para cada técnica foram elaborados modelos com parâmetros distintos amostrados de distribuições uniformes. Esses modelos foram gerados e analisados no `R` (versão 4.0.4) usando as bibliotecas `h2o` (versão 3.36.0.3) e `nlme` (3.1.155).

3.1 Métrica de erro personalizada

As aves mais ameaçadas, seja porque já possuem população pequena ou porque sofrem um elevado numero de capturas que pode reduzir sua população rapidamente no futuro, motivam a adoção de uma métrica de erro personalizada que permita escolher um modelo com maior capacidade de previsão nesses casos. A métrica de erro utilizada foi o erro médio quadrático ponderado por pesos específicos para cada linha de dado:

$$mse.w = mean(weight * (real - pred)^2)$$

Os pesos consistem na razão entre a quantidade de animais apreendidos e o tamanho de sua população para cada categoria de estado de conservação[2]. Desse modo, aves que são muito apreendidas ou que possuem população pequena terão pesos maiores no cálculo do erro.

4 Resultados

A seguir, os resultados são apresentados conforme a estratégia adotada.

4.1 Estratégia simplificada

Após o agrupamento dos dados nas covariáveis de interesse restaram 91 linhas, entretanto, com a adição de zeros, o numero de linhas passou para 135 (ou seja, 30% de casos com zero apreensão na série temporal completa). Foram elaborados 20 modelos nessa estratégia e os resultados de `mse.w` para a partição de teste estão apresentados na Tabela 1.

[2] à cada estado de conservação foi atribuído um valor numérico arbitrário na tentativa de refletir a ordem de grandeza da população.

Tabela 1: Valores de erro quadrático ponderado para 20 modelos de cada técnica de aprendizado na estratégia simplificada.

Técnica	Média	DesvPad	Mínimo
Média	3697	0	3697
GLM	2823	157	2595
RF	2739	72	2556
GBM	3038	110	2888
DL	2759	52	2645

Nota-se que a técnica de florestas aleatórias apresentou a melhor capacidade de predição, entretanto a raiz do erro quadrático médio é elevado (50.6). De fato, no gráfico da Figura 1 de valor predito contra valor real para o modelo RF cujo erro foi mínimo é possível notar que a previsão não é razoavelmente precisa.

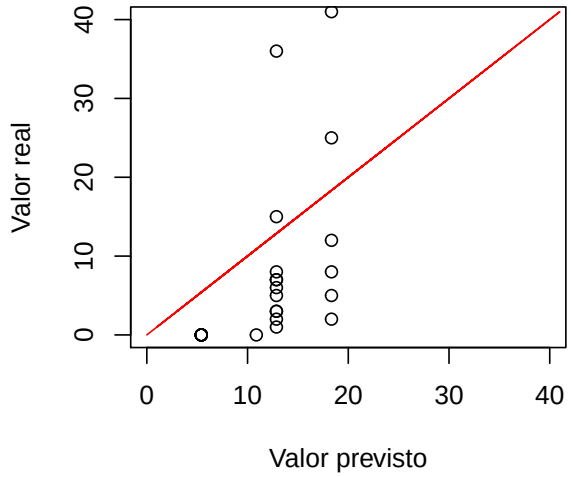


Figura 1: Valor previsto pelo melhor modelo na estratégia simplificada vs. valor real da quantidade de aves apreendidas

4.2 Estratégia intermediária

Após o agrupamento dos dados nas covariáveis de interesse restaram 7058 linhas, das quais os trios Latitude x Longitude x Qtde apreendida específicos das aves ameaçadas são mostrados na Figura 2.

Foram elaborados 20 modelos nessa estratégia e os resultados de `mse.w` para a partição de teste estão apresentados na Tabela 2. Novamente, observa-se que a técnica de florestas aleatórias apresentou a melhor capacidade de predição, entretanto a raiz do erro quadrático médio teve notável redução (20.6). Dos gráficos da Figura 3 de valor predito contra valor real para o modelo RF cujo erro foi mínimo é possível notar que a previsão ainda não é razoavelmente precisa.

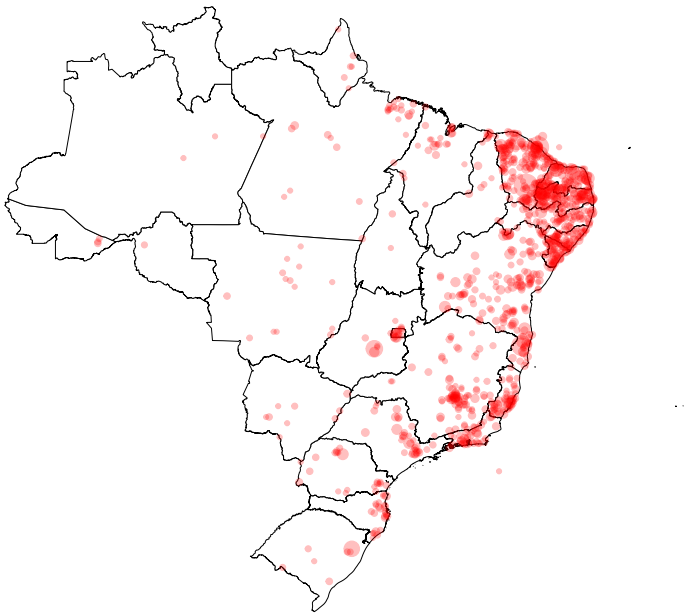
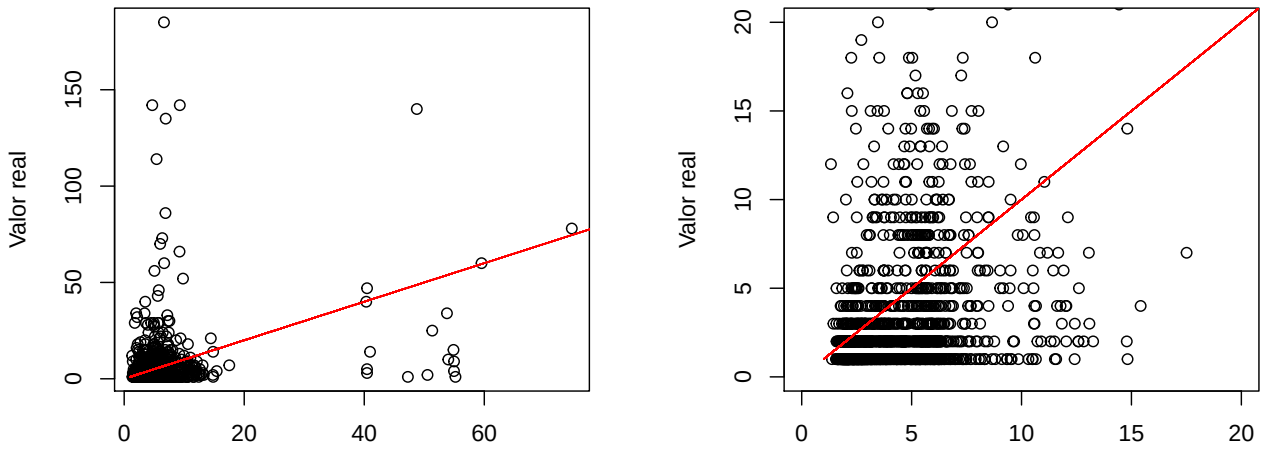


Figura 2: Mapa brasileiro da quantidade de aves ameaçadas apreendidas de 2014 a 2018.

Tabela 2: Valores de erro quadrático ponderado para 20 modelos de cada técnica de aprendizado na estratégia intermediária.

Técnica	Média	DesvPad	Mínimo
Média	437	0	437
GLM	429	1	426
RF	427	3	423
GBM	464	32	426
DL	430	2	427



(a) Intervalo completo

(b) Intervalo [0,20]

Figura 3: Valor previsto pelo melhor modelo na estratégia intermediária vs. valor real da quantidade de aves apreendidas.

4.3 Estratégia refinada

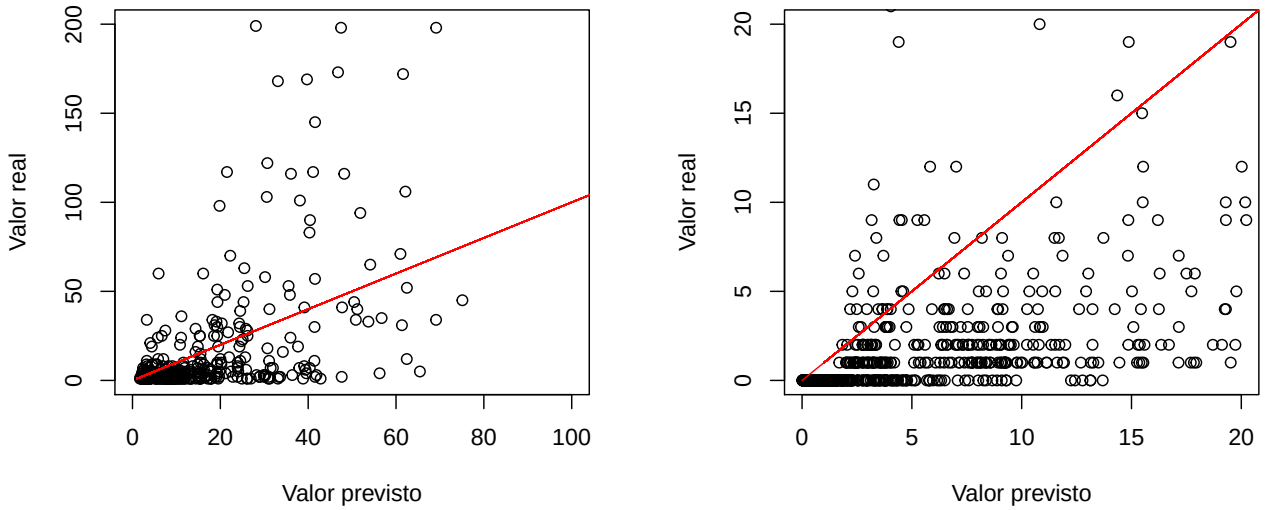
Após o agrupamento dos dados nas covariáveis de interesse restaram 1993 linhas, entretanto, com a adição de zeros, o numero de linhas passou para 5365 (ou seja, 63% de casos com zero apreensão na série temporal completa). Foram elaborados 20 modelos nessa estratégia e os resultados de `mse.w` para a partição de teste estão apresentados na Tabela 3.

Nessa estratégia, a técnica GLM obteve o menor `mse.w`, porém os valores médios e desvio padrão não indicam uma robustez na capacidade de previsão dessa técnica para esse problema nessa estratégia. A técnica com o mínimo `mse.w` em geral foi a florestas aleatórias, como encontrado nas estratégias anteriores. A raiz do erro quadrático médio é elevada (48.7).

Dos gráficos da Figura 4 de valor predito contra valor real para o modelo RF cujo erro foi mínimo é possível notar que a previsão ainda não é razoavelmente precisa.

Tabela 3: Valores de erro quadrático ponderado para 20 modelos de cada técnica de aprendizado na estratégia refinada.

Técnica	Média	DesvPad	Mínimo
Média	3188	0	3188
GLM	2506	394	2251
RF	2375	82	2304
GBM	4688	2498	2632
DL	2668	292	2297



(a) Intervalo completo

(b) Intervalo [0,20]

Figura 4: Valor previsto pelo melhor modelo na estratégia refinada vs. valor real da quantidade de aves apreendidas.

5 Discussão

Foi feita uma modelagem exploratória, buscando um conhecimento preliminar do potencial do dataset e da aplicação de algumas técnicas de aprendizado de máquina na solução do problema. Foi possível observar que, no geral, as técnicas de aprendizado tiveram um desempenho preditivo ligeiramente superior que o modelo de média estratificada por localização geográfica. A estratégia com melhor desempenho *não* contemplou a série temporal completa, o que é irreal, pois implicaria saber quais locais ocorrerão as apreensões futuras, cabendo ao modelo apenas a previsão da quantidade. Ainda assim o desempenho do modelo foi baixo. Além disso, pode ser mais importante prever onde será cometida a infração que a quantidade no ato da apreensão, o que sugere repensar a modelagem para obter como saída do modelo uma probabilidade de ocorrência de infração no local.

O baixo poder preditivo dos modelos sugere que as covariáveis selecionadas não contribuem para reduzir a variância suficientemente. De fato, é razoável supor que existem outras fontes de variância provenientes, por exemplo, de questões culturais locais, realidade socioeconômica, efetivo de fiscalização disponível, fatores de perturbação do habitat natural que desloquem as aves para regiões de maior concentração população, entre outras.

Outro fator de dificuldade na aplicação de modelos de previsão é a natureza ativa do agente que pratica a infração, de modo que locais que ocorrem mais apreensões podem ficar evidentes e deslocar o agente para outro local. Isso sugere que é válido buscar técnicas em outros campos da estatística, como da estatística Bayesiana ou de Teoria dos jogos que tentem contornar problemas dessa natureza.

Não foi feita pesquisa bibliográfica prévia sobre o problema e o uso de aprendizado de máquina nesse contexto, sendo essa tarefa um bom próximo passo em direção à melhoria da modelagem para solução do problema, tanto na área de ecologia e conservação da avifauna para identificar covariáveis de influência como na área estatística para observar como problemas dessa natureza são abordados.