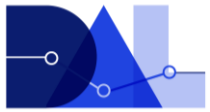


BRAZILIAN E-SCIENCE WORKSHOP

DESAFIOS NA PREDIÇÃO DO CONSUMO DE PESTICIDAS EM ESCALA GLOBAL USANDO APRENDIZADO DE MÁQUINA

Bruna Capistrano[1] , Luma Chen[1] , Matheus Ribeiro[1] , Carla Pacheco[1,3] , Dacy Lobosco[1] , João Quadros[1] , Maria Izabel Barreto[2] , Eduardo Ogasawara[1]

[1] CEFET/RJ – Centro Federal de Educação Tecnológica Celso Suckow da Fonseca. *Data Analytics Lab*
[2] PBIO - Petrobras Biocombustível
[3] PUC-Rio – Pontifícia Universidade Católica

Roteiro

- Motivação em monitoração de consumo de pesticidas
- Importância da predição do consumo de pesticidas pelos países
- Características e Desafios
- Definição e Tratamento do problema
- *State of art*
- Etapas da predição de séries temporais
- Análise dos Resultados
- Conclusões

Motivação

em monitoração de consumo de pesticidas

Importância

Redução de perdas de produtos agrícolas

Melhoria na qualidade e segurança alimentares

Controle do consumo de pesticidas pelas **demandas dos países**

Transição para a cultivo orgânico e **preservação do meio ambiente**

Problemas

Contaminação de plantas não-alvo

Poluição do meio-ambiente

Impactos negativos na saúde humana

Uso desenfreado dessas **substâncias tóxicas**

Definição de pesticidas

"substâncias ou mistura de substâncias para (...) mitigar qualquer peste, além de servirem como reguladores de plantas, desfolhantes ou dessecantes"

[Lee and Choi, 2020]

Impactos
no meio ambiente
(contaminação de solo e
de **águas do subsolo** e
resíduos nos alimentos)

Impactos na **saúde**
(intoxicação)

[Gomes et al., 2020]

Importância
da predição do consumo de pesticidas pelos países

Políticas ambientais

Planejamento agrícola

Redução do consumo de tais substâncias tóxicas

Características e Desafios

Dados de consumo de pesticidas em formato de séries temporais dos TOP 10 países consumidores mundiais (dados reais de [FAO, 2022])

Cenário extremo de *small-data* [Kitchin and Lauriault, 2015] (dados anuais e recentes)

Definição do problema

Dado o cenário de *small data*,
será que conseguimos usar métodos de

Aprendizado de Máquina (AM)

para melhorar

a predição

do consumo de pesticidas ?

Definição e Tratamento
do problema

Dadas as séries temporais dos TOP 10 países consumidores de pesticidas ...

... predizer o consumo com modelos de Aprendizado de Máquina (AM)...

... versus

ARIMA (modelo básico de referência), usando a estratégia:
combinar AM com Métodos de pré-processamento (MP).

Trabalhos Relacionados e *string de busca*

String de busca na base de dados Scopus em 04 de junho de 2022

("predict*" OR "forecast*") AND ("pesticide*") AND ("con- sumption" OR "demand" OR "usage") AND ("machine learning" OR "ARIMA")

... 29 artigos retornados,
sendo poucos sobre ST
e nenhum com referência direta
a avaliação de métodos preditivos para
o consumo de pesticidas em escala global
ou de países ...

... trabalhos mais aderentes ao tema:
Yu et al. [2020] com dados de níveis de resíduos de pesticidas em vegetais folhosos e amiláceos por 15 meses para predição de curto prazo dos níveis de resíduos usando ARIMA.

e Rao et al. [2021] que desenvolveu um modelo de CNN para prever praga ou doença que infecta as colheitas.

Referencial Teórico em

Séries Temporais (ST)

ST é qualquer sequência de observações de um fenômeno através do tempo.

Comumente, as STs são expressas a partir das suas componentes de **tendência**, **sazonalidade** e **ruído aleatório**
[Box et al., 2015].

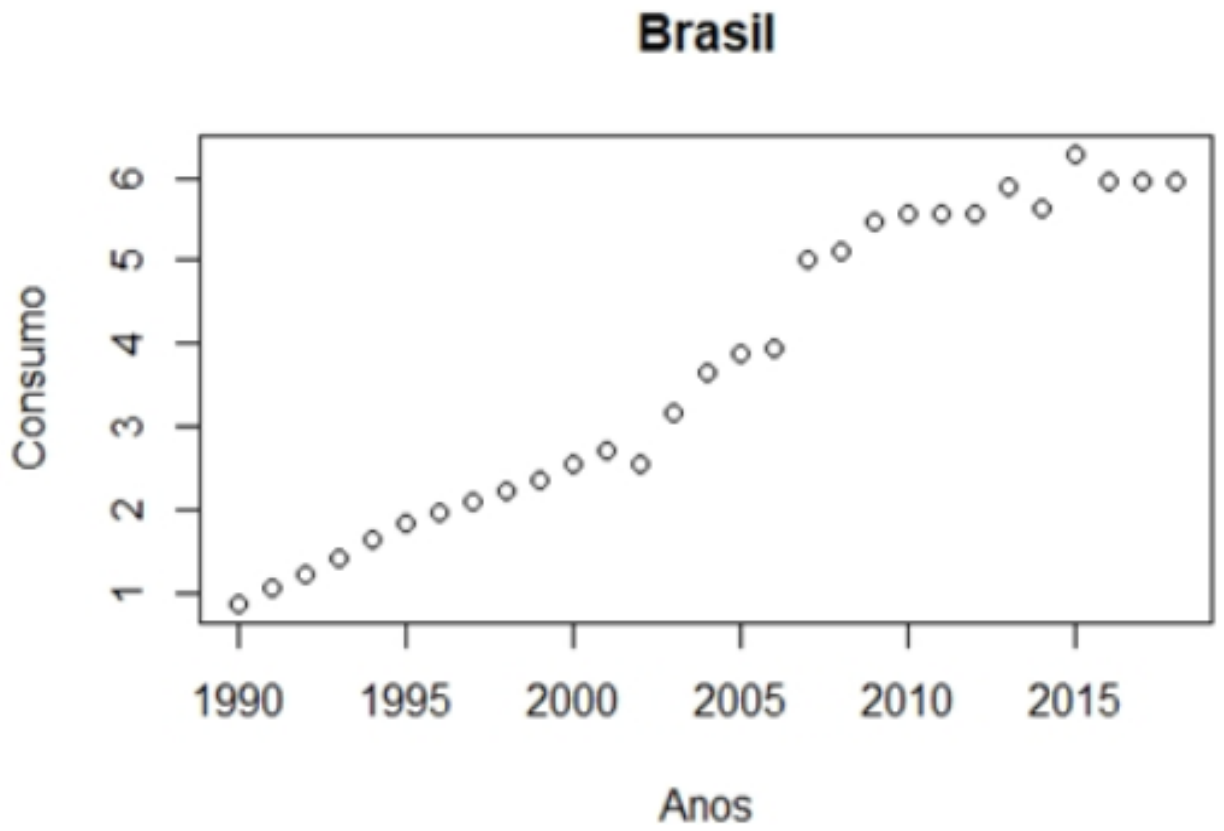


Fig.1: ST do consumo de pesticidas (kg/hectare) no Brasil, de 1990 a 2018

Tais propriedades não são constantes em várias aplicações reais, como a não estacionariedade.

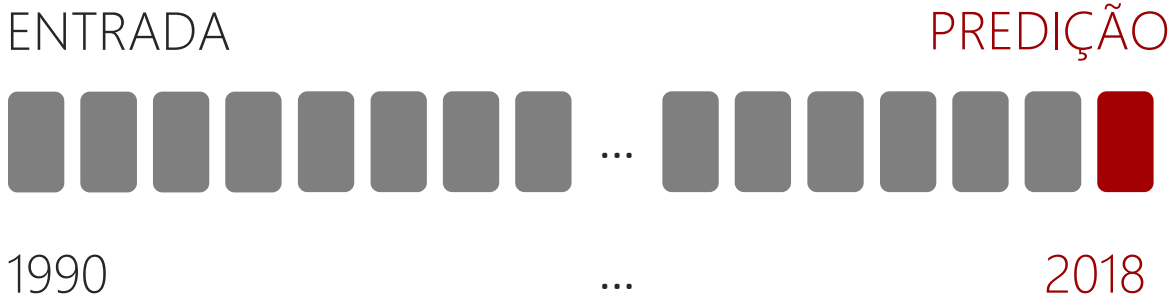
Predição de ST
em duas etapas

i) pré-processamento da entrada por normalização:
- janela deslizante
- global
- global com diferenciação
- adaptativa (AN)

ii) modelos:
- AM (6)
- ~~ARIMA (1)~~

Algoritmos de AM:

- Florestas Aleatórias de regressão (RF)
- Máquinas de Vetores de Suporte para regressão (SVR)
- Redes Neurais Artificiais
  - do tipo Perceptron de Multicamada (MLP)
  - Máquinas de Aprendizado Extremo (ELM)
  - Convolucionais (CNN)
  - Long-Short Term Memory (LSTM)



modelo linear versátil que pode apresentar limitações nos problemas que contenham padrões temporais não-lineares [Júnior et al., 2019]

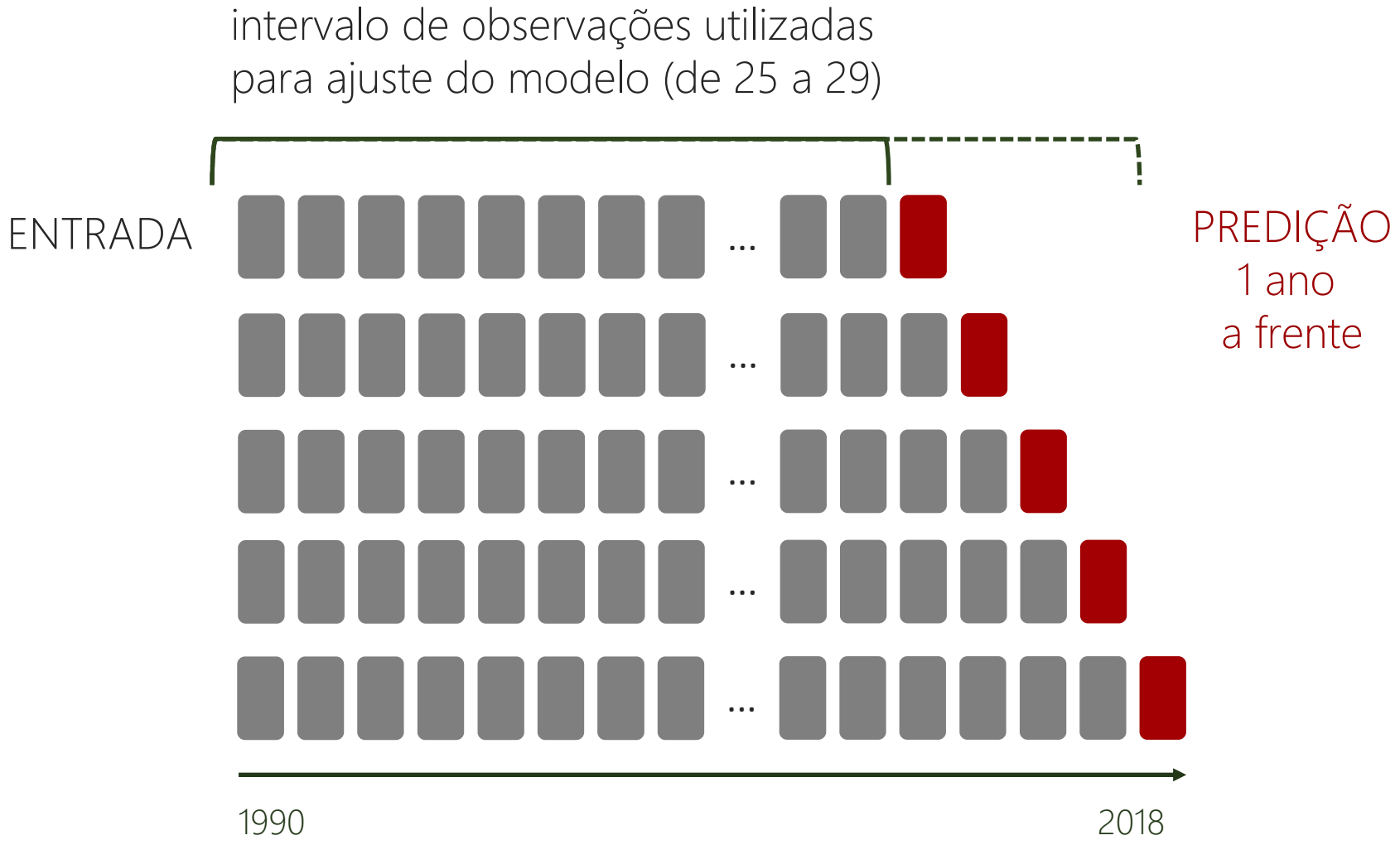
Validação Cruzada

em séries temporais

(*rolling origin*)

[Hyndman e Athanasopoulos, 2018]

- 1200 configurações
- 50 cenários experimentais



- otimização de hiperparâmetros para cada par MP e AM
  - tamanho da janela (de 4 a 5)
  - parâmetros específicos para os AM

Análise dos Resultados

diversidade de comportamento

séries crescentes
(como o caso do Brasil)

séries decrescentes
(como o caso do Japão)

séries alternadas
(como o caso da Turquia)

Análise dos Resultados

detalhamento I

erro de treinamento e erro de previsão, em SMAPE (%)

Buscamos o modelo que minimiza o erro médio simétrico

Os resultados mostram a média e o desvio padrão

Tabela 1. Erro em SMAPE(%) na predição por país no treino (△) e teste (□)

Países	ARIMA △	AM △	model	ARIMA □	AM □
Alemanha	$5,3 \pm 0,7$	$3,4 \pm 0,1$	RF + gmm	$3,8 \pm 2,9$	$5,1 \pm 1,4$
Brazil	$4,3 \pm 0,2$	$2,4 \pm 0,3$	RF + an	$6,9 \pm 2,2$	$6,5 \pm 3,6$
China	$2,2 \pm 0,1$	$1,3 \pm 0,1$	RF + an	$1,6 \pm 1,4$	$1,1 \pm 1,3$
EUA	$3,1 \pm 0,2$	$1,5 \pm 0,1$	RF + gmm	$0,3 \pm 0,2$	$0,9 \pm 1,1$
Franca	$7,7 \pm 0,2$	$4,7 \pm 0,6$	RF + an	$11,1 \pm 6,6$	$11,9 \pm 8,9$
India	$13,3 \pm 0,3$	$6,7 \pm 1,5$	SVR + gmm	$11,9 \pm 9,7$	$15,7 \pm 10,2$
Japao	$2,8 \pm 0$	$1,3 \pm 0,2$	SVR + gmm	$2,5 \pm 1,8$	$1,5 \pm 1,7$
Mexico	$10,2 \pm 0,3$	$5,8 \pm 1,3$	RF + swmm	$5,2 \pm 3,4$	$5,4 \pm 4,6$
Russia	$2,3 \pm 1,3$	$1,4 \pm 0,5$	SVR + gmm	$8,8 \pm 11,5$	$23 \pm 19,6$
Turquia	$11,8 \pm 0,3$	$6,6 \pm 1,6$	CNN + an	$14 \pm 10,9$	$9,5 \pm 9,2$

Melhores resultados dos algoritmos de AM, com a escolha do melhor par AM com MP identificado no treino

Melhor par AM com MP por país, identificado no treino

AM com MP no teste

ARIMA no teste

No geral, o ARIMA obteve os melhores resultados no teste

Análise dos Resultados

detalhamento II
cenário de teste
considerando todos os países

desempenho geral do ARIMA em 6.6% ± 7.3%

parâmetro *d* ajustado para 1

dos 50 modelos ajustados, 33 eram passeios aleatórios (com ou sem *drift*).

Tabela 2. Erro em SMAPE(%) nas predições dos pares AM+MP na etapa de teste

AM	gmm	swmm	an	gmmd
CNN	$15 \pm 15,5$	$6,9 \pm 7,2$	$9,5 \pm 9,1$	$9,5 \pm 9,1$
ELM	$7,3 \pm 8,7$	$10,6 \pm 12,3$	$7,8 \pm 9,2$	$8,7 \pm 10,4$
LSTM	$8,8 \pm 10,3$	$6,8 \pm 7,5$	$7,7 \pm 8,1$	$6,7 \pm 7,5$
MLP	$6,8 \pm 8,3$	$13,2 \pm 29,2$	$6,6 \pm 6,8$	$7,4 \pm 7,9$
RF	$7,4 \pm 9,1$	$11,3 \pm 13,4$	$8,1 \pm 8,2$	$7,4 \pm 7,3$
SVR	$9,5 \pm 11,5$	$10 \pm 11,3$	$7,5 \pm 8,1$	$7,5 \pm 8,1$

LSTM + *gmmd*

gmmd e *an* se destacaram por apoiar modelos baseados em choques

MLP + *an*

Fazem frente ao ARIMA essas duas combinações de AM + MP no teste
(esses modelos não foram os melhores no treino)

→ a predição da próxima observação é um *choque* em relação à observação anterior.

não captura padrões relevantes em relação aos termos defasados.

Obs: o *choque* caracteriza o valor aleatório em relação ao quanto a série oscila.

Conclusões

Neste cenário de *small data*, os MP são importantes na predição (os resultados são limitados devido a pouca quantidade de dados)

Trabalhos futuros:
combinar MP com outras técnicas, como *amostragem* e *preparação de dados,* que levem mais em consideração os *termos mais recentes* para que os AM consigam modelar melhor os choques.

o ARIMA não trouxe um ganho de conhecimento em termos dos processos que regem as séries de pesticidas

SIMPÓSIO BRASILEIRO DE BANCO DE DADOS

DESAFIOS NA PREDIÇÃO DO CONSUMO DE PESTICIDAS EM ESCALA GLOBAL USANDO APRENDIZADO DE MÁQUINA

Bruna Capistrano[1] , Luma Chen[1] , Matheus Ribeiro[1] , Carla Pacheco[1,3] ,
Dacy Lobosco[1] , João Quadros[1] , Maria Izabel Barreto[2] , Eduardo Ogasawara[1]
[1] CEFET/RJ – Centro Federal de Educação Tecnológica Celso Suckow da Fonseca. *Data Analytics Lab*
[2] PBIO - Petrobras Biocombustível
[3] PUC-Rio – Pontifícia Universidade Católica