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[3], Dacy Lobosco[1], João Quadros[1], Maria Izabel Barreto[2], Eduardo Ogasawara[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

[2]Petrobras Biocombustível - PBIO

[3]Pontifícia Universidade Católica - PUC-Rio

{bruna.capistrano,luma.chen,matheus.ribeiro}@aluno.cefet-rj.br

cpacheco@inf.puc-rio.br, {dacy.lobosco,joao.quadros}@cefet-rj.br

maizabel@petrobras.com.br, eogasawara@ieee.org

Resumo. *O consumo de pesticidas é relevante para o agronegócio, governo e sociedade em escala mundial. Tal consumo é um insumo fundamental na cadeia produtiva de alimentos, sendo um indicador importante para o monitoramento dos níveis de intoxicações e de resíduos que degradam o meio ambiente. Analisar o consumo de pesticidas em escala global ao longo do tempo é um grande desafio, pois os dados disponíveis são anuais e recentes. Este trabalho explora diferentes maneiras de otimizar a construção de modelos de previsão, utilizando diferentes abordagens por meio de combinações pareadas entre pré-processamento de dados e métodos de aprendizado de máquina. Foram avaliadas essas abordagens para obter previsões baseadas em dados reais sobre pesticidas nos dez principais países que os consomem. Os resultados mostraram que a utilização de modelos de aprendizado de máquina com desempenho satisfatório é difícil de se obter, considerando esse cenário de poucos dados e, ao mesmo tempo, peculiar de acordo com o país.*

Abstract. *The consumption of pesticides is relevant to agribusiness, government, and society on a global scale. Such consumption is a fundamental input in the food production chain and an important indicator for monitoring the levels of poisoning and residues that degrade the environment. Analyzing pesticide consumption on a global scale over time is a major challenge, as the available data are annual and recent. This work explores different ways to optimize the construction of prediction models, using different approaches through paired combinations between data pre-processing and machine learning methods. These approaches were evaluated to obtain predictions based on real data of pesticides in the top ten countries that consume them. The results showed that using machine learning models with satisfactory performance is difficult to obtain, considering this scenario of very small data and, at the same time, peculiar according to the country.*

1. Introdução

A demanda por alimentos deve crescer em 50% até 2050, devido ao aumento da população mundial [FAO, 2022]. Sustentar tal produção é um considerável desafio, principalmente, sem comprometer a integridade do meio ambiente. Há o consenso de que se alcançar esta demanda de produção agrícola é essencial para a política global, estabilidade e equidade sociais [Tilman et al., 2002]. Embora o uso de pesticidas auxilie no aumento de produção, a utilização de materiais tóxicos pode prejudicar o solo, os alimentos, a saúde humana e a de animais.

Os pesticidas são quaisquer substâncias ou mistura de substâncias que tenham como objetivo prevenir, destruir, repelir ou mitigar qualquer peste, além de servirem como reguladores de plantas, desfolhantes ou dessecantes [Lee and Choi, 2020]. Tais substâncias podem ser orgânicas ou inorgânicas, podendo ser aplicadas na plantação, nas sementes ou no solo, por meio de borrifadores, fumaça ou em pó. O uso de pesticidas pode contaminar as águas do subsolo, deixar resíduos nos alimentos, além de intoxicar agricultores que aplicam diretamente essas substâncias químicas [Gomes et al., 2020].

Por conta disso, considera-se que a agricultura precisa adaptar, procurar e expandir a adoção de práticas sustentáveis para mitigar os efeitos no clima, saúde e uso do solo [FAO, 2022]. Tais mitigações podem ser alcançadas quando apoiadas pela monitoração e predição do consumo de pesticidas. A predição permite um acompanhamento geral da quantidade demandada por país, evitando o uso desenfreado de tais substâncias tóxicas. Esse acompanhamento permite, por exemplo, traçar estratégias de transição para o cultivo orgânico de alimentos. Isso contribui para a saúde da população, preservação do meio ambiente, a qualidade e segurança alimentares, que são pontos críticos da economia e indústria mundiais [Gomes et al., 2020].

Os trabalhos relacionados concentram-se na predição do consumo de pesticidas em cenários de plantações locais [Yu et al., 2020]. A contribuição deste trabalho está na análise da predição do consumo de pesticidas em escala global por meio da combinação de Métodos de Pré-processamento (MP) e de Métodos de Aprendizado de Máquina (AM) em um cenário de muitos poucos dados. Esta combinação é um diferencial deste trabalho e foi explorada em dados reais de séries temporais dos dez maiores países consumidores de pesticidas do mundo. Foram testadas 1.200 configurações para todos os modelos em todos esses países, cada um deles passando por otimização de hiperparâmetros. Essas configurações foram comparadas como o modelo ARIMA [Box et al., 2015], adotado como um modelo básico de referência.

O artigo está organizado em mais cinco seções. As Seções 2 e 3 descrevem conceitos gerais de predição de séries temporais e os principais trabalhos relacionados. A Seção 4 detalha a metodologia utilizada neste trabalho, enquanto a Seção 5 apresenta a avaliação experimental e uma análise mais detalhada sobre predição do consumo de pesticidas. Finalmente, a Seção 6 apresenta a conclusão e trabalhos futuros.

2. Predição em séries temporais

Uma série temporal (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]. Na prática, pode ser observado que tais

propriedades não são constantes em várias aplicações reais, tais como séries envolvendo fenômenos socioeconômicos [Tsay, 2010], que caracterizam a não estacionariedade.

A predição de STs é feita em duas etapas: (i) pré-processamento da entrada e (ii) utilização de modelos (estatísticos ou de aprendizado de máquina) para predizer observações futuras baseadas nas observações dadas como entrada. Em (i), as técnicas de normalização são aplicadas para utilização de AM [Esling and Agon, 2012]. Incluem-se na lista: normalização de cada janela deslizante (padroniza cada janela entre 0 e 1), normalização global de todas as janelas (padroniza a matriz de todas as janelas entre 0 e 1), normalização global de todas as janelas após diferenciação (o mesmo caso anterior, após aplicação do *backshift*) e normalização adaptativa (AN) [Salles et al., 2019]. As duas primeiras técnicas são mais sensíveis a cenários de não-estacionariedade.

Alguns métodos estatísticos são utilizados como base para predição de ST. O ARIMA (modelo Autorregressivo Integrado de Média Móvel) é um dos métodos mais usados para análise de dados em ST, em consequência de sua generalidade [Box et al., 2015]. Ele contém três componentes: Autoregressivo (AR), filtro de Integração (I) e Médias Móveis (MA) [Box et al., 2015]. Por estar na categoria de modelos lineares, o ARIMA pode apresentar limitações nos problemas que contenham padrões temporais não-lineares [Júnior et al., 2019], mas são bastante versáteis e possuem diversos métodos para otimização deste modelo [Hyndman and Athanasopoulos, 2018].

Na família dos métodos de AM, há inúmeros métodos, onde se podem destacar as Florestas Aleatórias de regressão (RF), as Máquinas de Vetores de Suporte para regressão (SVR), Redes Neurais Artificiais do tipo Perceptron de Multicamada (MLP) [Han et al., 2012; Esling and Agon, 2012], as redes de Máquinas de Aprendizado Extremo (ELM) [Huang et al., 2012], que possuem uma única camada intermediária e aprendem os pesos sem iteração. No que concerne às redes profundas, destacam-se as Convolucionais (CNN) e as *Long-Short Term Memory* (LSTM) que possuem recursos adicionais para memorizar a sequência de dados [Siami-Namini et al., 2019] e suas múltiplas dependências de tempo [Lindemann et al., 2021]).

3. Trabalhos Relacionados

Devido a relevância socioeconômica do uso de pesticidas, diversos artigos foram publicados no tema. Neste trabalho, foi elaborada a seguinte *string* de busca sobre predição do consumo de pesticidas: (“predict*” OR “forecast*”) AND (“pesticide*”) AND (“consumption” OR “demand” OR “usage”) AND (“machine learning” OR “ARIMA”), na base de dados *Scopus*. O termo “forecast*” da *string* de busca já contempla implicitamente o termo “séries temporais”. Vale ressaltar que esta pesquisa foi realizada em 04 de junho de 2022, quando foi encontrado um total de vinte e nove artigos, sendo poucos sobre ST.

Neste cenário, os trabalhos retornados mais aderentes ao tema em questão são de Yu et al. [2020] e de Rao et al. [2021]. O primeiro trabalho apresenta um conjunto de dados dos níveis de resíduos de pesticidas em vegetais folhosos e amiláceos (fontes de amido) medidos por quinze meses. Foram feitos testes de amostras pareadas para construção de quatro modelos ARIMA de ST. Os modelos ARIMA foram eficazes para a predição de curto prazo dos níveis de resíduos de pesticidas agrícolas. Tais modelos podem ser usados preventivamente, como alertas de potencial contaminação, com base

em padrões reais de uso, tipo de cultura, estação, além de outros parâmetros.

O segundo trabalho desenvolveu um modelo de CNN para prever praga ou doença que infecta as colheitas. Esse e outros trabalhos mostraram-se um pouco mais distantes nesta área de pesticidas. Nenhum artigo que fizesse referência direta a avaliação de métodos preditivos para o consumo de pesticidas em escala global ou de países foi encontrado.

4. Predição de pesticidas

Esta seção explora maneiras de otimizar a construção de modelos para predição do consumo de pesticidas, em escala global, através de diferentes combinações pareadas entre MP e AM. Os MP avaliados neste trabalho foram: *normalização por janelas deslizantes* (swmm), *normalização global das janelas* (gmm), *normalização global das janelas após diferenciação* (gmmd) e *normalização adaptativa* (an). Os MP foram combinados com os seguintes AM: ELM, MLP, RF, SVR, CNN e LSTM. Adicionalmente, o modelo ARIMA foi usado como base de comparação. Nenhuma técnica de *data augmentation* ou *transfer learning* foram utilizadas.

Tais métodos foram aplicados aos dados de séries temporais reais dos dez maiores países consumidores de pesticidas (Brasil, Japão, China, Estados Unidos, França, Índia, Alemanha, México, Rússia e Turquia). Ao todo, foram utilizadas dez séries com 29 observações correspondentes aos anos de 1990 a 2018, obtidas livremente em [FAO, 2022]. Considerando-se o número de observações disponível, o trabalho está inserido no cenário bem extremo de *small-data* [Kitchin and Lauriault, 2015].

Os cenários experimentais foram definidos a partir das dez séries temporais e a variação do intervalo de observações utilizadas (de 25 a 29), para estabelecer um cenário de validação cruzada de séries temporais (*i.e.*, *rolling origin*) [Hyndman and Athanasopoulos, 2018] com predições de um passo à frente. Este procedimento visa garantir a qualidade dos resultados. Ao todo, foram 50 cenários experimentais que sofreram otimização de hiperparâmetros para cada par MP e AM, além do ARIMA. Foram otimizados alguns hiperparâmetros, incluindo o tamanho da janela (de 4 a 5) e hiperparâmetros específicos para os AM descritos a seguir.

No ELM e MLP, variou-se de 3 a 8 neurônios na camada escondida. Para ELM, fixou-se a função de ativação *purelin* e para o MLP, variou-se o *decay* entre 0 e 1 em intervalos de 0.01 (parâmetro de regularização para evitar superajuste). No RF, foi variado o tamanho das florestas entre 10 e 50. No SVR, o *epsilon* variou entre 0 e 1 em intervalos de 0.05 e a função de custo da margem de 1 a 20, em intervalos de 1. Nas CNN e LSTM, variou-se de 3, 5, 8, 16 e 32 neurônios na camada escondida e foram estabelecidas 1.000 épocas.

A solução foi desenvolvida em R [Shumway and Stoffer, 2017]. Esses modelos foram construídos em paralelo em um servidor Intel i7-10700 com 2.90GHz, 16 núcleos, 128GB de RAM usando Ubuntu 20.04 LTS. Após a previsão, o erro de treinamento e o erro de previsão, em SMAPE (%), de cada modelo foram armazenados e sumarizados para comparação (medindo a média e o desvio padrão). Tal métrica ajuda a identificar o modelo que minimiza o erro médio simétrico.

5. Análise dos Resultados

O consumo de pesticidas varia de forma diferente ao longo do tempo, podendo aumentar (como foi o caso do Brasil), decrescer (como no Japão) e alternar entre esses padrões (como na Turquia). Por conta da diversidade de comportamento, tanto o ARIMA quanto os modelos AM, oscilaram bastante no desempenho de predição, conforme mostra a Tabela 1. Quando se seleciona o melhor par de AM com MP na fase de treino, observa-se um melhor resultado em relação ao ARIMA, como no caso de China, Japão e Turquia. Entretanto, na fase de testes, no geral, o ARIMA foi superior.

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	$14 \pm 10,9$	$9,5 \pm 9,2$

Para se aprofundar a análise, a Tabela 2 mostra o cruzamento do desempenho dos métodos AM com os MP na fase de testes, considerando-se todos os países. Vale ressaltar que o desempenho geral do ARIMA, nesta situação, ficou em $6.6\% \pm 7.3\%$. Observa-se que, no geral, poucas combinações de AM + MP são capazes de fazer frente ao ARIMA, destacando-se MLP + an e LSTM + gmmd.

Três considerações devem ser destacadas. A primeira refere-se aos modelos ARIMA. Em todos, o parâmetro d foi ajustado para 1. Dos 50 modelos ajustados, 33 eram passeios aleatórios (com ou sem *drift*). Dos demais, alguns tinham p ou q igual a 1. Isso deixa claro que a predição da próxima observação é um choque em relação à observação anterior, não capturando padrões relevantes em relação aos termos defasados. O choque caracteriza o valor aleatório em relação ao quanto a série oscila. A segunda é que os modelos de AM que tiveram melhor desempenho no teste não foram os melhores no treino. Em terceiro, deve ser apontado que tanto gmmd quanto an foram os MP que se destacaram por conseguirem apoiar modelos baseados em choques.

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	$9,5 \pm 11,5$	$10 \pm 11,3$	$7,5 \pm 8,1$	$7,5 \pm 8,1$

6. Conclusão

Este trabalho abre o espaço para predições de consumo de pesticidas em escala mundial apoiado por AM. Neste cenário de *small data*, ficou claro que os métodos de pré-

processamento são importantes no processo de predição, mas que precisam ser combinados 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. Os resultados são limitados em relação a pouca quantidade de dados. Adicionalmente, ficou claro que o ARIMA, apesar de ter apresentado melhor resultado na maioria dos casos, não trouxe um ganho de conhecimento em termos dos processos que regem as séries de pesticidas.

Referências

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Esling, P. and Agon, C. (2012). Time-series data mining. *ACM Computing Surveys*, 45(1).

FAO (2022). Food and Agriculture Organization of the United Nations. Technical report, http://www.fao.org.

Gomes, H. d. O., Menezes, J., da Costa, J., Coutinho, H., Teixeira, R., and do Nascimento, R. (2020). A socio-environmental perspective on pesticide use and food production. *Ecotoxicology and Environmental Safety*, 197.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*.

Huang, G.-B., Zhou, H., Ding, X., and Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(2):513–529.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Júnior, D. d. O. S., Oliveira, J. d., and Neto, P. d. M. (2019). An intelligent hybridization of ARIMA with machine learning models for time series forecasting. *Knowledge-Based Systems*, 175:72–86.

Kitchin, R. and Lauriault, T. (2015). Small data in the era of big data. *GeoJournal*, 80(4):463–475.

Lee, G.-H. and Choi, K.-C. (2020). Adverse effects of pesticides on the functions of immune system. *Comparative Biochemistry and Physiology Part - C: Toxicology and Pharmacology*, 235.

Lindemann, B., Müller, T., Vietz, H., Jazdi, N., and Weyrich, M. (2021). A survey on long short-term memory networks for time series prediction. In *Procedia CIRP*, volume 99, pages 650–655.

Rao, C., Rahul, M., Dasgupta, S., and Hegde, R. (2021). Sustainable Pesticide usage in Agriculture. In *2021 IEEE Mysore Sub Section International Conference, MysuruCon 2021*, pages 628–633.

Salles, R., Belloze, K., Porto, F., Gonzalez, P., and Ogasawara, E. (2019). Nonstationary time series transformation methods: An experimental review. *Knowledge-Based Systems*, 164:274–291.

Shumway, R. H. and Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples*. Springer, New York, NY, 4 edition.

Siami-Namini, S., Tavakoli, N., and Siami Namin, A. (2019). A Comparison of ARIMA and LSTM in Forecasting Time Series. In *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, pages 1394–1401.

Tilman, D., Cassman, K., Matson, P., Naylor, R., and Polasky, S. (2002). Agricultural sustainability and intensive production practices. *Nature*, 418(6898):671–677.

Tsay, R. S. (2010). *Analysis of Financial Time Series*. Wiley, Cambridge, Mass, 3 edition.

Yu, W., Han, X., Wang, Y., and Yang, J. (2020). Prediction of pesticide residues in agricultural products based on time series model in Chengdu, China. In *IOP Conference Series: Earth and Environmental Science*, volume 594.