



REDES NEURAIS ARTIFICIAIS PARA PREVISÃO DE OCORRÊNCIA DE PSILÍDEO NO CINTURÃO CITRÍCOLA BRASILEIRO

João V. P. **de Souza**¹; Marina R. M. **Monteiro**²; Yan Y. **Luz**³; Sônia **Ternes**⁴; Maria F. **Moura**⁵;
Alécio S. **Moreira**⁶; Kleber X. S. **de Souza**⁷

Nº 25601

Resumo: *Uma das doenças mais destrutivas dos citros no mundo, é o Huanglongbing (HLB), tendo como vetor o psilídeo Diaphorina citri, que vem afetando drasticamente a produção de diversos países. Em um cenário onde o Brasil está posicionado como maior produtor de laranja do mundo, identificar as regiões que mais favorecem a ocorrência do psilídeo é de suma importância para garantir a produção de futuras plantações. Com base em dados climáticos e geográficos, buscamos criar um modelo de rede neural que possa prever a ocupação desses insetos em determinadas regiões, mais especificamente no cinturão citrícola brasileiro.*

Palavras-chave: Citros, HLB, rede neural, inteligência artificial

¹ Autor, Bolsista CNPq (PIBIC): Graduação em Estatística, UNICAMP, Campinas-SP;
joaovitor.souza@colaborador.embrapa.br

² Bolsista Fapesp Treinamento Técnico IV: Mestre em Matemática Aplicada e Computacional, UNICAMP, Campinas-SP;
marina.monteiro@colaborador.embrapa.br

³ Bolsista Fapesp Iniciação Científica: Graduação em Ciência de Dados e Inteligência Artificial, PUC-Campinas, Campinas-SP; yan.luz@colaborador.embrapa.br

⁴ Pesquisadora da Embrapa Agricultura Digital, Campinas-SP; sonia.ternes@embrapa.br

⁵ Pesquisadora da Embrapa Agricultura Digital, Campinas-SP; maria-fernanda.moura@embrapa.br

⁶ Pesquisador da Embrapa Agricultura Digital, Campinas-SP; alecio.moreira@embrapa.br

⁷ Orientador: Pesquisador da Embrapa Agricultura Digital, Campinas-SP; kleber.sampaio@embrapa.br



Abstract: *One of the most destructive citrus diseases in the world is Huanglongbing (HLB), whose vector is the psyllid *Diaphorina citri*, which has drastically affected production in several countries. Given Brazil's position as the world's leading orange producer, identifying the regions that most favor the occurrence of the psyllid is of significant importance to ensure future plantations. Based on climatic and geographical data, we aim to create a neural network model that can predict the occupation of these insects in certain regions, specifically in the Brazilian citrus belt.*

Keywords: Citrus, HLB, neural network, artificial intelligence

1. INTRODUÇÃO

Entre o período de 2024-2025, o Brasil foi responsável por 70% da produção mundial de suco de laranja e 29% da produção de laranja (Estados Unidos, 2025), ocupando a posição de maior produtor mundial dos dois produtos. Este é um dos ramos da fruticultura de maior relevância no mundo, principalmente por seus benefícios para a saúde de seus consumidores. Seus frutos são fontes de vitamina C, auxiliam no fortalecimento da imunidade, ajudam na cicatrização de feridas e queimaduras, e têm um importante papel na prevenção da coagulação sanguínea (Costa, 2020).

Com posição de destaque global na citricultura, atividade agrícola que se refere ao cultivo de frutas cítricas, como laranja, tangerina e limão, diversos desafios são encontrados no Brasil. Um dos maiores problemas globais atuais neste ramo (Castilhos *et al*, 2017) é a contaminação das plantações por bactérias associadas ao *Huanglongbing* (HLB), uma das doenças mais destrutivas dos citros. Esta bactéria tem como principal vetor o psíldeo *Diaphorina citri* (Albuquerque *et al*, 2017), no qual já foram desenvolvidos alguns métodos de proteção contra este psíldeo, como o uso periódico de inseticidas e diferentes métodos de manejo. Apesar disto, por meio do vetor, esta doença vem se espalhando pelo Brasil, especialmente no cinturão citrícola brasileiro, principal região do país na produção de citros, que engloba os estados de Minas Gerais e São Paulo.

Saber a quantidade desses psíldeos presentes nas plantações é de suma importância para identificarmos padrões de ocorrências, como período anual de maior reprodução, e assim agir mais assertivamente na prevenção da doença. Ao longo dos anos foram coletados dados de diversas fazendas pertencentes ao cinturão citrícola brasileiro, como dados climáticos e quantidade quinzenal de psíldeos capturados nas armadilhas de cada talhão.

Com base nessas variáveis de predição e informações climáticas futuras, podemos criar um modelo de rede neural que estima a porcentagem de ocorrência do psíldeo em determinada região e período do ano. Com isso, utilizarmos desta informação para escolhermos de maneira estratégica em quais locais plantar novas culturas de citros e nas plantações já cultivadas, podemos fortalecer medidas de prevenção antes de momentos críticos.

2. MATERIAL E MÉTODOS

A base de dados para treinar este modelo de rede neural é composta por informações coletadas das fazendas por meio de uma parceria com a Fundecitrus (Fundo de Defesa da Citricultura), como ocupação do inseto (quantidade de psilídeos), manejo utilizado, área da fazenda, entre outros. Os dados de clima foram coletados por meio do *Climate Hazards Center InfraRed Precipitation with Station data* (CHIRPS), FieldClimate e NasaPower.

Uma das principais tarefas para conseguirmos um bom resultado com o modelo é a preparação da base de dados. Assim nos próximos subtópicos são apresentadas as técnicas que foram utilizadas para lidar com variáveis não numéricas, dados faltantes e intervalos muito distintos.

2.1. Pré-processamento

Foram agrupados os dados das fazendas, enviados pela Fundecitrus, com os de clima das estações meteorológicas. Com isso, a base de dados original continha 322074 linhas e 150 colunas. Para que houvesse um melhor desempenho no treinamento do modelo, foram abordadas diversas técnicas de limpeza dos dados, descritas nos subtópicos a seguir.

2.2.1. Agrupamento

Primeiramente, como estamos lidando com um modelo de regressão, é necessário que existam apenas valores numéricos na base de dados.

O código foi desenvolvido com base na escala municipal, juntando todas as linhas de um mesmo município. Como pode ser visto na **Figura 1**, foi feita uma média simples ao agrupar os dados com base nas colunas com o código IBGE do município e a data de registro.

```
# Selecionando apenas colunas com valores numericos
numerical_columns = df.drop(columns=['tempo', 'CD_MUN']).select_dtypes(include=[np.number]).columns

# Calcula media simples
df_mun = df.groupby(['tempo', 'CD_MUN'])[numerical_columns].mean().reset_index()
```

Figura 1. Agrupamento por município

Além da escala municipal, foram abordadas outras estratégias de agrupamento. Ao juntarmos os dados em mesorregiões agrupamos cada vez mais o problema em busca de uma maior precisão.

Como pode ser consultado na **Tabela 1**, os dados também foram agrupados em grupos de 4 e 6 mesorregiões vizinhas, todos com média simples. Entretanto, é necessário um limite geográfico de agrupamento para que o resultado final não seja generalista demais e não consigamos utilizar este resultado na prática.

Tabela 1. Descrição das escalas de agrupamentos utilizadas.

Escala	Descrição
Municipal	Municípios do cinturão citrícola
Mesorregião_Original	Mesorregiões definidas pelo IBGE
Mesorregião_4	Mesorregiões originais agrupadas em 4
Mesorregião_6	Mesorregiões originais agrupadas em 6

2.2.2. Melhores variáveis preditoras

De todas as variáveis utilizadas inicialmente, foi percebido que os modelos aprenderam melhor com menos variáveis preditoras, então foram escolhidas as 30 melhores de acordo com o resultado do treinamento. Na **Tabela 2** é possível observar dois sufixos em comum entre algumas colunas que foram utilizadas.

Tabela 2. Descrição dos sufixos utilizados nos nomes das variáveis

Sufixo	Descrição
_15	Intervalo dos 15 dias anteriores à data do registro
_d[1-8]	<i>Delays</i> quinzenais. Ex: d1 = 15 dias; d4 = 60 dias

As variáveis que melhor performaram são, em geral, relacionadas ao clima, como temperatura média, máxima e mínima e ao tamanho das plantações de citros, como pode ser observado na **Tabela 3**.

Tabela 3. Descrição das variáveis utilizadas no treinamento do modelo.

Variável	Descrição
Ano	Ano de registro
Área Citros	Área de citros
Índice Área Citros	Área de citros dividida pela área do município
Razão Total	Área de citros dividida pela quantidade de propriedades de citros
T_15_[d4, d5, d6, d7, d8]	Média da temperatura média
Tabsmx_15	Maior valor registrado de temperatura máxima

Tabsmax_15_[d1, d6, d7, d8]	Maior valor registrado de temperatura máxima
Tabsmin_15_[d1, d3, d4, d5, d6, d7, d8]	Menor valor registrado de temperatura mínima
Trange_15	Tabsmax - Tabsmin
Trange_15_[d1, d2, d3, d4, d5, d8]	Tabsmax - Tabsmin
prec_15_d3	Precipitação em mm
balanco_h_cad100_d3	Balanço Hídrico com Capacidade de Água Disponível de 100mm

2.2.3. Manejos

A coluna 'tipo_manejo' pode receber os valores A, B, C ou D, e refere-se ao rigor com que o psilídeo (*Diaphorina Citri*) e a doença são controlados. Em geral, do manejo A em direção ao D há menos rigor no controle.

Como o manejo C e D apresentam um baixo controle da doença, espera-se que a ocorrência seja maior nestes casos, podendo resultar em alguma espécie de *outlier* na variável de ocupação de psilídeos, que é o nosso alvo. Então, foram testados alguns modelos sem esses 2 manejos para observar se existe alguma melhora aparente no resultado final.

Entretanto, devido à pequena quantidade de dados de manejo C e D quando comparados aos manejos A e B, não encontramos uma diferença significativa ao testarmos o modelo excluindo os manejos C e D. Por isso, no modelo final foram usados todos os quatro tipos de manejos.

2.2.4. Normalização

Dentre as diversas variáveis contidas na base de dados, algumas apresentam grandes intervalos, como por exemplo a temperatura média que vai de 12°C a 33°C, quando comparada com o balanço hídrico que vai de -140mm até 501mm. Em regressão, uma grande variabilidade nos intervalos dos dados pode dificultar muito o aprendizado do modelo. Para contornar esse problema, uma técnica de normalização foi utilizada.

Além disso, com a normalização dos dados, o modelo tende a ter uma melhora na generalização, aplicando de forma mais precisa o que foi aprendido para valores não identificados anteriormente. A rede lida com todas as variáveis tendo o mesmo peso inicial, facilitando na convergência e estabilidade numérica para atingirmos um bom resultado durante o treinamento do modelo.

Foi utilizado a classe *StandardScaler* da biblioteca *SciKit-Learn* (SCIKIT-LEARN, 2025) como método de padronização dos valores numéricos. Cada coluna é transformada para que tenha média igual a 0 e desvio padrão igual a 1, com base na **Equação 1**.

$$z = \frac{x - \underline{u}}{S} \quad (1)$$

Onde, x é o valor original, $\underline{u}$ é a média da distribuição e S é o desvio padrão da distribuição; z é o valor padronizado.

2.3. Modelo Utilizado

Redes neurais são modelos computacionais de inteligência artificial, inspiradas no funcionamento do cérebro humano para reconhecer padrões. Sua composição é fundamentalmente feita por diversas camadas com “neurônios artificiais” conectados entre si. Dentre as principais camadas estão uma de entrada dos dados, uma de saída que vai mostrar o resultado final e uma ou mais camadas ocultas, onde os pesos são ajustados para melhorar o desempenho final. É nesta última parte que ocorrem as principais diferenças entre os modelos de redes neurais.

Este modelo foi utilizado com o objetivo de identificar padrões complexos na base de dados, uma vez que estamos lidando principalmente com variáveis relacionadas ao clima e buscando encontrar algum padrão na ocorrência do inseto vetor (*Diaphorina Citri*).

2.3.1. FFN (*FeedForward Network*)

Para a execução deste projeto, foi utilizado o modelo de rede neural FFN (*FeedForward Network*), ou seja, uma rede de avanço, onde os dados seguem um fluxo para a frente (sem ciclos), diferente de, por exemplo, redes neurais recorrentes. Por meio de algoritmos como *backpropagation* e gradiente descendente, a rede neural ajusta os pesos e vieses das conexões até passar pela camada de saída.

O modelo definido contém a seguinte estrutura: uma camada de entrada com 64 neurônios totalmente conectados e a função de ativação *ReLU*, uma camada oculta com 32 neurônios e também a função *ReLU*, por fim uma camada de saída com apenas 1 neurônio, pois estamos prevendo um valor numérico, sem nenhuma função de ativação.

2.4. Validação

As métricas utilizadas para avaliar o desempenho do modelo foram: R^2 , responsável por mostrar o quanto o modelo explica a variância dos dados; MAE, média das diferenças absolutas entre os valores reais e previstos; e RMSE, métrica utilizada para penalizar mais os grandes erros.

Além disso, foi utilizada a curva de perda para avaliar se o modelo está convergindo e detectar se existe algum *overfitting*. Com base nas métricas numéricas e no gráfico, as camadas foram sendo adaptadas, adicionando ou diminuindo a quantidade de camadas ocultas e neurônios na estrutura da rede neural.

3. RESULTADOS E DISCUSSÃO

Os modelos foram treinados para cada método de agrupamento diferente, utilizando o valor de 100 *epochs* como base, seus resultados podem ser consultados na **Tabela 4**.

Tabela 4. Resultados dos modelos utilizando diferentes métodos de agrupamento.

Agrupamento	R ²	MAE	RMSE
Municipal	0.27	0.19	0.25
Mesorregião	0.16	0.18	0.21
Mesorregião_4	0.33	0.15	0.19
Mesorregião_6	0.34	0.16	0.20

Os melhores modelos que obtivemos foram resultantes do agrupamento por 4 e 6 mesorregiões, onde os valores das 3 métricas que utilizamos (R², MAE e RMSE) são semelhantes. Ambos ficam em destaque quando comparados com os resultados de outros métodos de agrupamento, como o municipal.

Podemos observar que a escala municipal explica menos a variância dos dados (R² = 0.27) e tem um erro médio um pouco maior (MAE = 0.19 e RMSE = 0.25) em relação às mesorregiões.

Na **Figura 2**, é possível comparar a diferença entre a curva de perda durante o treinamento do modelo (com *validation_split* = 0.2) e quando aplicada nos dados de validação. Como ambas as curvas estão próximas, é um sinal de que o modelo está aprendendo sem causar um sobreajuste, onde não apenas memoriza os dados de treino mas também consegue generalizar para dados até então não observados. Além disso, a curva diminui suavemente até se estabilizar em torno de 50 *epochs*, mostrando que a partir deste ponto não houve melhorias significativas.

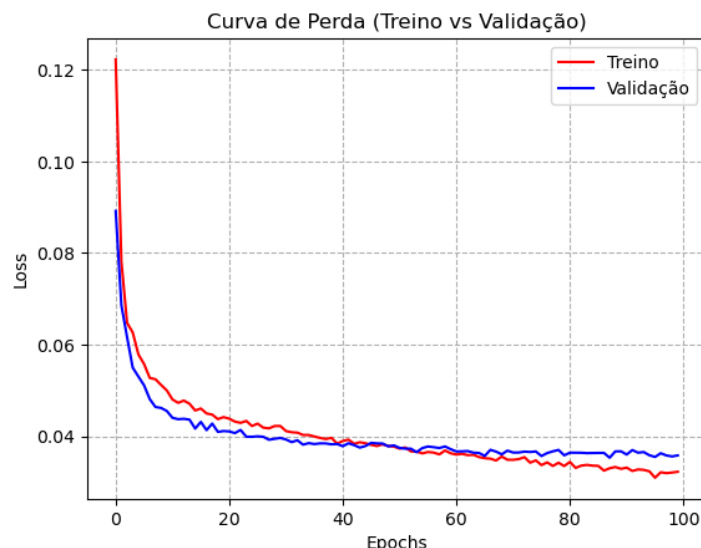


Figura 2. Curva de perda do modelo com escala de 6 mesorregiões.

Para gerar a curva de perda, foi utilizado um *validation_split* = 0,2 durante o treinamento. Este resultado foi obtido após a complexidade do modelo ter sido reduzida para menos camadas ocultas e com menos neurônios, direcionando o modelo a uma melhor generalização e convergência, tornando-o mais equilibrado com relação ao viés e variância.

4. CONCLUSÃO

Neste trabalho utilizamos um dos modelos fundamentais de rede neural (FFN) com o intuito de prever a porcentagem de ocupação do inseto vetor (*Diaphorina Citri*) de uma das doenças mais destrutivas dos citros, o HLB (Albuquerque 2017).

O modelo composto por apenas 1 camada oculta, foi treinado com base em diferentes tipos de agrupamentos, podendo prever a ocupação municipal ou em diferentes escalas de mesorregiões. Tendo como melhor resultado o agrupamento por 6 mesorregiões, com $R^2 = 0,34$.

Apesar deste resultado não ser satisfatório para implementação em uma aplicação prática, foi possível conhecer melhor o problema e alguns caminhos a serem seguidos para atingirmos um ótimo resultado.

O objetivo é atingir um R^2 maior que 0.9, para tornar-se uma previsão mais precisa, e reduzir o MAE e RMSE para o mais próximo de 0. Para isso podemos utilizar a mesma base de dados com outros algoritmos de inteligência artificial.

Modelos mais complexos, como Redes Neurais Recorrentes (RNNs), podem identificar melhor padrões difíceis de serem encontrados, o que pode ter um bom desempenho para o nosso problema, mas também pode causar muito sobreajuste. Já modelos mais simples, como *Random Forests* (RFs), também podem ser um ótimo caminho a se seguir, no qual identificando os padrões mais simples podemos conseguir um bom resultado.



5. AGRADECIMENTOS

Este trabalho foi desenvolvido no âmbito do projeto Fapesp 2023/02915-8 com apoio do Fundo de Defesa da Citricultura (Fundecitrus). A bolsa do primeiro autor foi financiada pelo programa PIBIC/CNPq.

6. REFERÊNCIAS

ALBUQUERQUE, T. de F.; FANCELLI, M.; SOUSA, M. R. Aspectos comportamentais de *Diaphorina citri* KUWAYAMA e *Tamarixia radiata* (WATERSTON) frente a voláteis de *Murraya paniculata* e *Ruta graveolens*. In: JORNADA CIENTÍFICA EMBRAPA MANDIOCA E FRUTICULTURA, 17., 2017, Cruz das Almas. **Ciência e Empreendedorismo**: resumos. Cruz das Almas: Embrapa Mandioca e Fruticultura, 2017. 1 p. Disponível em: <https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1083716/1/MDPA409517V02Aprovado.pdf>. Acesso em: 2 jun. 2025.

CASTILHOS, R. V.; BRUGNARA, E. C.; MARO, L. A. C.; CANALE, M. C. Huanglongbing: possível ameaça à citricultura catarinense? **Agropecuária Catarinense**, v. 30, n. 1, p. 36-39, jan./abr. 2017. DOI: <https://doi.org/10.52945/rac.v30i1.37>.

COSTA, J. B. **A Importância dos Citros**. 2020. Disponível em: <https://codeagro.agricultura.sp.gov.br/cesans/artigo/222/a-importancia-dos-citros>. Acesso em: 2 jun. 2025.

ESTADOS UNIDOS. Department of Agriculture. Foreign Agricultural Service. **Top producing countries: 2024/2025 oranges production**. Washington, DC, [2025]. Disponível em: <https://www.fas.usda.gov/data/production/commodity/0571120>. Acesso em: 2 jun. 2025.

SCIKIT-LEARN. **StandardScaler**. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>. Acesso em: 2 jun. 2025.