UNIVERSIDADE ESTADUAL DE PONTA GROSSA

SETOR DE CIÊNCIAS AGRÁRIAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

JOÃO PAULO MENDES DE ALMEIDA

MÉTODOS CLÁSSICOS E BASEADOS EM APRENDIZADO DE MÁQUINA PARA PREVISÃO DE PREÇO DE TOMATE *IN NATURA*

PONTA GROSSA

2023

JOÃO PAULO MENDES DE ALMEIDA

MÉTODOS CLÁSSICOS E BASEADOS EM APRENDIZADO DE MÁQUINA PARA PREVISÃO DE PREÇO DE TOMATE *IN NATURA*

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada, curso de Mestrado em Computação Aplicada da Universidade Estadual de Ponta Grossa, como requisito parcial para obtenção do título de Mestre.
Orientação: Profª. Drª. Alaine Margarete Guimarães
Coorientação: Prof. Dr. Eduardo Fávero Caires

PONTA GROSSA

2023

Almeida, João Paulo Mendes de
A447 Métodos clássicos e baseados em aprendizado de máquina para previsão de preço de tomate in natura / João Paulo Mendes de Almeida. Ponta Grossa, 2023.
94 f.

Dissertação (Mestrado em Computação Aplicada - Área de Concentração: Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.

Orientadora: Profa. Dra. Alaine Margarete Guimarães.
Coorientador: Prof. Dr. Eduardo Fávero Caires.

1. Agricultura. 2. Mercado atacadista. 3. Variáveis preditoras. 4. Análise de variáveis. I. Guimarães, Alaine Margarete. II. Caires, Eduardo Fávero. III. Universidade Estadual de Ponta Grossa. Computação para Tecnologias em Agricultura. IV.T.

CDD: 004

Ficha catalográfica elaborada por Maria Luzia Fernandes Bertholino dos Santos- CRB9/986

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
Av. General Carlos Cavalcanti, 4748 - Bairro Uvaranas - CEP 84030-900 - Ponta Grossa - PR - https://uepg.br

TERMO

TERMO DE APROVAÇÃO

João Paulo Mendes de Almeida

MÉTODOS CLÁSSICOS E BASEADOS EM APRENDIZADO DE MÁQUINA PARA PREVISÃO DE PRECO DE TOMATE IN NATURA

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Profª. Drª. Alaine Margarete Guimarães (UEPG - Presidente)

Prof. Dr. Arion de Campos Júnior (UEPG)

Prof. Dr. Tiago Pellini (IDR-PR)

Ponta Grossa, 28 de setembro de 2023.

Documento assinado eletronicamente por **Alaine Margarete Guimaraes**, **Professor(a)**, em 28/09/2023, às 16:21, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Arion de Campos Junior**, **Professor(a)**, em 28/09/2023, às 16:21, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **TIAGO PELLINI**, **Usuário Externo**, em 27/11/2023, às 15:34, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida no site https://sei.uepg.br/autenticidade informando o código verificador **1615251** e o código CRC **C0055ABD**.

AGRADECIMENTOS

À Universidade Estadual de Ponta Grossa, que me acolheu em seu ambiente acadêmico. Agradeço a todos os professores, funcionários e colegas de curso que contribuíram para o meu crescimento acadêmico e profissional.

À minha orientadora, Alaine Margarete Guimarães, pelo apoio, orientação precisa e paciência demonstrada ao longo de todo o processo de pesquisa. Suas valiosas orientações foram cruciais para o desenvolvimento deste trabalho.

À minha família, pelo apoio emocional e incentivo constante. Vocês foram a minha base durante os desafios desta jornada acadêmica, e por isso sou imensamente grato.

Aos professores e pesquisadores que generosamente compartilharam seu conhecimento e colaboraram com esta pesquisa.

Finalmente, gostaria de expressar minha gratidão a todos que, de alguma forma, contribuíram para esta jornada. Obrigado por fazerem parte deste capítulo da minha vida.

Muito obrigado a todos.

RESUMO

Prever os preços agrícolas é crucial para a tomada de decisões e impacta a renda dos agricultores, os preços dos alimentos e a economia como um todo. O tomate é uma das principais olerícolas produzidas e comercializadas no Brasil, com um mercado dinâmico afetado por fatores como variações climáticas e sazonalidade. Este estudo tem como objetivo comparar o desempenho de diferentes métodos clássicos e baseados em aprendizado de máquina na previsão dos preços do tomate nos mercados atacadistas Ceagesp e Ceasa/PR. Os métodos de previsão ARIMA, SARIMA, ARIMAX, SVR, LSTM e CNN foram utilizados, e dados de séries temporais sobre os preços do tomate e outras variáveis correlacionadas foram coletados entre 2010 e 2021. O teste ADF foi utilizado para determinar a ordem de diferenciação necessária para tornar a série estacionária. A análise de componentes principais foi usada para reduzir a dimensionalidade dos dados, extraindo um número menor de componentes que representam a maior parte da variação observada nos dados. O teste de Ljung-Box foi aplicado para analisar os resíduos e verificar a adequação do modelo aos dados observados. O desempenho dos diferentes modelos foi analisado utilizando as métricas RMSE e MAPE. A análise da importância das variáveis é uma etapa importante nos modelos de aprendizado de máquina e foi aplicada para identificar as variáveis preditoras mais influentes sobre a variável de resposta. O teste estatístico não paramétrico de Wilcoxon foi usado para avaliar a diferença entre os modelos candidatos. Os resultados mostraram que algumas variáveis estavam altamente correlacionadas entre si. Este estudo concluiu que o modelo SVR teve uma precisão maior em comparação com outros modelos. A escolha do método mais adequado pode variar de acordo com o objetivo da previsão, o período, a disponibilidade e a qualidade dos dados, entre outros fatores.

Plavras-chave: agricultura, mercado atacadista, variáveis preditoras, análise de variáveis.

ABSTRACT

Predicting agricultural prices is crucial for decision-making in the agribusiness sector, impacting farmers' income, food prices, and the economy as a whole. Tomatoes are one of the main vegetables produced and traded in Brazil, with a dynamic market affected by factors such as climatic variations and seasonality. This study aims to compare the performance of different classical and machine learning-based methods in predicting tomato prices in the Ceagesp and Ceasa/PR wholesale markets. The ARIMA, SARIMA, ARIMAX, SVR, LSTM, and CNN forecasting methods were used, and time series data on tomato prices and other correlated variables were collected between 2010 and 2021. The ADF test was used to determine the order of difference required to make the series stationary. Principal component analysis was used to reduce data dimensionality by extracting a smaller number of components that represent most of the variation observed in the data. The Ljung-Box test was applied to analyze residuals and verify model adequacy to observed data. The performance of different models was analyzed using RMSE and MAPE metrics. Variable importance analysis is an important step in machine learning models and was applied to identify the most influential predictor variables on the response variable. The non-parametric Wilcoxon statistical test was used to evaluate the difference between candidate models. Results showed that some variables were highly correlated with each other. This study concluded that the SVR model had higher accuracy compared to other models. The choice of the most suitable method may vary according to the objective of the forecast, period, availability, and quality of data, among other factors.

Keywords: agriculture, wholesale market, predictor variables, variable analysis.

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE SIGLAS

ACP – Análise de Componentes Principais

AIC – Critério de Informação de Akaike

ANP – Agência Nacional do Petróleo

ARIMA - *Autoregressive Integrated Moving Average*

AttLSTM - *Attention-based Long Short-Term Memory*

BPNN - *Back Propagation Neural Network*

Ceagesp – Companhia de Entrepostos e Armazéns Gerais de São Paulo

Ceasa/PR - Central de Abastecimento do Paraná

Cepea - Centro de Estudos Avançados em Economia Aplicada

Conab – Companhia Nacional de Abastecimento

Deral – Departamento de Economia Rural do Paraná

DNN - *Deep Neural Network*

ETS - *Exponential Smoothing*

GBDT - *Gradient boosting decision tree*

GRNN - *Generalized Neural Network*

GRU - *Gated Recurrent Unit*

IBGE – Instituto Brasileiro de Geografia e Estatística

LSTM - *Long Short-Term Memory*

MAE - *Mean Absolute Error*

MAPE - *Mean Absolute Percentage Error*

ML – *Machine Learning*

MSE - *Mean Squared Error*

POF – Pesquisa de Orçamento Familiar

RBF - *Radial Basis Function Network*

RF - *Random Forest*

RL – Regressão Linear

RMSE - *Root Mean Squared Error*

RNA – *Artificial Neural Network*

RNN - *Recurrent Neural Network*

SARIMA – *Seasonal Autoregressive Integrated Moving Average*

STL - *Seasonal-trend Decomposition Using Loess*

SVR - *Support Vector Machine*

SUMÁRIO

1 INTRODUÇÃO

A predição de preços agrícolas é uma ferramenta importante para auxiliar produtores, comerciantes e governos na tomada de decisão no setor agropecuário. Os preços podem influenciar a renda dos agricultores, os custos dos alimentos e a economia como um todo. Dessa forma, a capacidade de prever com precisão os preços futuros é essencial para planejar a produção e os investimentos.

Para Rao (1989) a lei da oferta e demanda é fundamental para a economia. Quando a primeira é maior do que a segunda, os preços tendem a cair, enquanto que, quando o inverso ocorre, os preços tendem a subir. Essa lei é aplicável a todas as cadeias de suprimentos, incluindo a cadeia agroalimentar.

De acordo com a FAO (2013), os produtos agrícolas são essenciais para o consumo diário. Porém, o excesso pode fazer com que os preços caiam, prejudicando os agricultores, enquanto que a falta de eleva as cotações, o que pesa para os consumidores. É importante entender essas dinâmicas para desenvolver políticas e estratégias que promovam a estabilidade dos mercados.

Como proposto por Rao (1989), as condições climáticas e políticas governamentais também afetam a dinâmica da formação de preços. O clima está diretamente ligado às condições de estabelecimento de cultivos, e a ocorrência de eventos extremos pode prejudicar o lado da oferta. A política governamental interfere por meio de subsídios, tarifas e regulamentações.

A flutuação cambial é outro fator que afeta a formação de preços agrícolas, especialmente em países que dependem de exportações, como o Brasil. Esse fator interfere na competitividade dos produtos nos mercados internacionais e, consequentemente, afeta os preços domésticos.

O estado encontra dificuldades para controlar fatores que afetam os preços agrícolas e fornecer previsões mais precisas pode ajudar a reduzir riscos e contribuir para a estabilidade do mercado (SERRA; GIL, 2007).

O Brasil é um grande produtor de tomate, em 2021 foram colhidas, aproximadamente, 3,6 milhões de toneladas. Essa fruta ocupa o terceiro lugar em termos de valor de produção e

de comercialização no setor de produtos *in natura*, ficando atrás apenas da cebola e banana, respectivamente (IBGE, 2020; IBGE, 2021).

Assim como para os demais produtos agrícolas, o mercado de tomate no Brasil é dinâmico e pode ser afetado por diversos fatores, como variações climáticas e sazonalidade. Por isso, a previsão de preços do tomate *in natura* deve ser relvante par as partes interessadas.

De acordo com Conab (2021), essa empresa pública é responsável por realizar estudos de mercado e levantamentos de safras para fornecer informações sobre a oferta e demanda de produtos agrícolas, principalmente grãos. A Companhia utiliza diversas técnicas de modelagem matemática e estatística para realizar suas previsões, como a análise de séries temporais e a modelagem econométrica.

Para Shumway e Stoffer (2017) os métodos clássicos para previsão de séries temporais, como ARIMA, SARIMA e ARIMAX, são baseados em componentes autorregressivos e de médias móveis. Esses métodos têm uma base teórica sólida e propriedades estatísticas bem estabelecidas, e são amplamente utilizados na literatura. Além disso, são relativamente eficientes em termos computacionais, o que facilita o ajuste e a aplicação mesmo com grandes conjuntos de dados.

No entanto, os métodos clássicos para previsão de séries temporais têm certas limitações, como a necessidade de estacionariedade e linearidade. Em alguns casos, podem ser necessárias abordagens mais avançadas, como modelos não lineares baseados em redes neurais ou árvores de decisão, para capturar melhor a complexidade dos dados. (HYNDMAN; ATHANASOPOULOS, 2018).

O uso de técnicas de aprendizado de máquina e séries temporais multivariadas para predição de preços de produtos *in natura*, em destaque os de tomate, no mercado brasileiro, é um problema ainda não abordado no âmbito da pesquisa acadêmica, o que motiva a proposta deste estudo, que está organizado da seguinte forma: seção 3, apresentação dos objetivos; seção 4, revisão de literatura sobre o tema; seção 5, apresentação dos métodos e materiais que foram utilizados; seção 6, apresenta e discute os resultados alcançados; seção 7, principais conclusões sobre o que foi abordado; e referências bibliográficas.

2 OBJETIVOS

O objetivo principal desta dissertação de mestrado é realizar uma comparação entre os desempenhos de diferentes métodos clássicos e de aprendizado de máquina para a tarefa de previsão de séries temporais para os preços de tomate nos mercados atacadistas Ceagesp, na cidade de São Paulo, e Ceasa/PR, em Curitiba.

Para alcançar esse objetivo, foram coletados dados de séries temporais de preços de tomate e outras variáveis correlacionadas. Esses dados, então compilados e organizados em uma base, serviram de entrada para os modelos de previsão. Cada modelo foi treinado e testado para, ao fim, ter o desempenho avaliado.

3 REVISÃO DE LITERATURA

3.1 CARACTERIZAÇÃO DE SÉRIES TEMPORAIS

Para Haykin (2009) uma série temporal pode ser definida como uma sequência de observações ordenadas em intervalos de tempo iguais. Essas observações podem ser representadas como uma função matemática, onde o valor em um determinado tempo é a observação correspondente na série. Essa função é chamada de processo estocástico.

A Tabela 1 exemplifica uma série fictícia.

Tabela 1 - Série temporal fictícia para exemplificar o conceito de série temporal

Data	Valor
2019-01-01	100
2019-02-01	120
2019-03-01	135
2019-04-01	145
2019-05-01	130
2019-06-01	155
2019-07-01	170
2019-08-01	180
2019-09-01	160
2019-10-01	175
2019-11-01	200
2019-12-01	220

Fonte: O autor

No exemplo, é apresentada uma sequência temporal com os valores mensais de uma determinada variável ao longo do ano de 2019. Cada linha representa uma observação em um determinado mês. A coluna "Data" indica a data da observação e a coluna "Valor" indica o valor associado a ela. Essa série pode ser usada para analisar o comportamento dos valores ao longo do tempo, identificar possíveis padrões sazonais ou tendências.

Essas análises envolvem a aplicação de diversos métodos matemáticos para descrever e prever o comportamento da série (HYNDMAN; ATHANASOPOULOS, 2018).

A escolha do método adequado para a caracterização depende do comportamento dos dados e do objetivo da análise. Se a série apresenta sazonalidade, pode ser necessário utilizar aquele que captura essa condição e prevê o comportamento futuro.

Diversas ferramentas estatísticas e de programação são utilizadas na análise de séries temporais, como o software R e o pacote estatístico da linguagem de programação Python. Essas ferramentas possuem funções específicas para a aplicação de modelos numéricos, além de ferramentas gráficas para a visualização dos dados e resultados.

A compreensão da estacionariedade, tendência e sazonalidade de uma série temporal é importante para selecionar a modelagem adequada e realizar previsões precisas. Esses componentes podem variar em intensidade e importância, portanto, é essencial caracterizar cada um deles (HYNDMAN; ATHANASOPOULOS, 2018).

3.1.1 Estacionariedade

Segundo Brockwell e Davis (2016), a estacionariedade é uma propriedade importante em séries temporais, pois muitos métodos estatísticos assumem que os dados possuem essa condição. Essa propriedade ocorre quando a média e a covariância das observações não dependem do tempo, ou seja, são constantes ao longo de toda a observação realizada.

Para verificar se uma série temporal é estacionária, é comum utilizar o conceito de estacionariedade fraca ou estacionariedade em sentido amplo. De acordo com Shumway e Stoffer (2017), a condição fraca requer apenas que a média e a covariância sejam constantes ao longo do tempo, enquanto que a em sentido amplo requer que as distribuições conjuntas sejam invariantes ao longo do tempo.

Existem diversas técnicas para transformar uma série temporal não estacionária em estacionária, como a diferenciação de primeira ordem, a transformação logarítmica e a aplicação de métodos autorregressivos. Box; Jenkins (1976) mostram que a aplicação de diferenciação de primeira ordem é uma das técnicas mais comuns e efetivas para tornar uma série temporal estacionária.

Segundo Cryer e Chan (2008), em algumas situações é possível trabalhar com séries temporais não estacionárias, desde que sejam utilizadas técnicas apropriadas. Alguns métodos de aprendizado de máquina são competentes na análise de sequências de dados que não satisfazem a condição.

3.1.2 Tendência

Para Morettin e Toloi (2018) a análise de tendência é um procedimento que permite identificar a direção e a magnitude da inclinação de uma série temporal. A tendência pode ser crescente, decrescente ou constante. Essa característica pode ser modelada por meio de uma regressão linear, na qual a média da série temporal é estimada por uma reta (BOX; JENKINS, 1976).

Ao ajustar uma regressão para modelar a tendência de uma série temporal, é importante verificar se os resíduos são aleatórios. Isso significa que não deve haver padrões não capturados pelo modelo (SHUMWAY; STOFFER, 2017). Caso os resíduos não sejam aleatórios, pode ser necessário utilizar outras técnicas, como a diferenciação ou a modelagem de componentes sazonais (MORETTIN; TOLOI, 2018).

A análise da tendência também pode ser feita por meio de gráficos, como o gráfico de dispersão ou o gráfico de médias móveis (BOX; JENKINS, 1976). No gráfico de médias móveis, é possível suavizar a série temporal calculando o valor médio de um certo número de pontos próximos, o que permite melhor visualização (SHUMWAY; STOFFER, 2017).

3.1.3 Sazonalidade

A sazonalidade é um padrão recorrente que se repete em intervalos de tempo regulares em uma série temporal. Chatfield (2003) define a sazonalidade como a variação periódica no comportamento da série ao longo do tempo. Essas variações são, muitas vezes, associadas a eventos, como estações do ano, dias da semana, entre outros.

Uma maneira de identificar a presença de sazonalidade em uma série temporal é através do gráfico de decomposição, que mostra os diferentes componentes separadamente. De acordo com Montgomery et al. (2015), esse gráfico é útil para visualizar o padrão de variação ao longo do tempo.

Para Hyndman e Athanasopoulos (2018), a componente sazonal pode ser modelada por meio dos modelos de média móvel ou de diferenças sazonais.

Outra forma de modelar a sazonalidade é através do modelo multiplicativo sazonal. Segundo Box et al. (2015), esse modelo é mais apropriado quando a amplitude da componente sazonal varia com o nível da série.

3.2 SÉRIES TEMPORAIS MULTIVARIADAS

Uma série temporal multivariada tem mais de uma variável dependente do tempo. Cada variável depende não apenas de seus valores passados, mas também de outras séries. Essa interdependência é usada para previsão de valores futuros.

Em modelos de séries temporais multivariadas, o conjunto preditor inclui múltiplas sequências univariadas que podem contribuir para a previsão da variável alvo. A escolha dessas séries é um processo complexo que envolve a experiência e o conhecimento do domínio (STOCK, 2001).

Em muitos contextos, os conjuntos de dados analisados podem conter um grande número de variáveis, ou seja, possuir alta dimensionalidade. Essa condição pode causar problemas para métodos de aprendizado de máquina, como aumento do custo computacional, dificuldade de interpretação e *overfitting (*HASTIE et al., 2009*).*

À medida que o número de variáveis aumenta, os cálculos necessários para treinar os modelos se tornam mais complexos e demorados, o que exige recursos computacionais significativos e pode tornar o processo de treinamento lento e ineficiente. Com grande número de dimensões, torna-se mais desafiador representar graficamente o espaço de características e entender as relações entre elas.

O *overfitting* ocorre quando os métodos de aprendizado de máquina se ajustam demais aos dados de treinamento, o que acarreta a captura de ruídos e padrões aleatórios em vez de relações verdadeiramente significativas (MURPHY, 2012). Isso pode resultar em um modelo que não generaliza bem para novos dados e leva a previsões imprecisas.

3.3 FORMAÇÃO DE PREÇOS DE PRODUTOS *IN NATURA*

Compreender os sinais do mercado, como tendências de preço, pode facilitar a tomada de decisão e apontar oportunidades para extrair maior valor para as empresas e consumidores finais.

Do lado da oferta, condições climáticas, custo de transporte, custos de produção e a variação de preços pagos ao agricultor, tornam essa abordagem ainda mais complexa. Enquanto variáveis como poder de compra e atividade econômica influenciam a demanda por produtos.

Os fatores climáticos são mais desafiadores na produção de hortaliças do que para outras culturas, como grãos. Portanto, condições diferentes das ideais podem impactar a

bioquímica, anatomia, morfologia e fisiologia desse grupo de plantas (FAHAD et al. 2017). As condições do clima também influenciam diretamente na pressão de pragas e doenças (CHAUHAN et al. 2014).

Fatores climáticos também exercem influência sobre o comportamento dos consumidores. Em períodos de clima quente, as pessoas tendem a buscar alimentos mais leves e refrescantes, o que pode aumentar a demanda por saladas, sucos e pratos frios, com uso de tomates, Por outro lado, em climas mais frios, os consumidores podem optar por alimentos mais quentes e preparações cozidas, reduzindo a demanda por vegetais *in natura* (SOLOMON, 2022).

Os custos para produção de alimentos hortifrutigranjeiros é variável importante para entender a formação do preço ao consumidor e no atacado. Podem ser levados em consideração os desembolsos efetivos realizados pelo produtor durante o ciclo produtivo, o que engloba gastos com mão de obra, reparos, insumos e manutenção de máquinas e benfeitorias específicas, além do valor da depreciação. Desses, o item insumos, como fertilizantes, é o que mais impacta o custo operacional total (REZENDE et al., 2005).

No estudo de Rezende et al. (2015), os autores analisaram a relação entre os valores dos principais insumos e os preços de *commodities* no Brasil, no período de 1996 a 2012. Os resultados mostraram que as cotações dos fertilizantes exercem uma influência significativa na formação dos preços dos principais produtos agrícolas brasileiros, como soja, milho e algodão.

Outro fator com potencial para influenciar toda a cadeia de suprimentos de produtos *in natura* é a variação do preço dos combustíveis, em especial o diesel. A transmissão dos reajustes dos preços no custo de transporte por parte dos prestadores é automatica. Além disso, o aumento do preço do diesel impacta os valores das operações agrícolas, implicando maiores custos de produção (PÉRA et al., 2018).

A inflação para os consumidores em geral é medida por índices compostos por diversos produtos e serviços, distribuídos em classes. Desse conjunto de itens, há aqueles que, apesar de comprometerem parcela relevante do orçamento familiar, demonstram comportamento pouco previsível, como as hortaliças e legumes, que apresentam comportamento alternado entre os alimentos que depositam contribuições positivas e negativas à formação dos índices (BRAZ, 2005).

O valor pago aos agricultores é uma variável que, em uma análise preliminar, pode interferir na formação dos preços. Uma referência que aborda a relevância do preço pago aos agricultores para o valor final de produtos agrícolas é o artigo de Souza e Viana (2007). Eles

analisaram a evolução histórica dos preços pagos aos agricultores para os principais produtos do Estado do Rio Grande do Sul.

Para Magalhães (2011) não há um padrão claro de precedência temporal dos preços de *commodities* agrícolas em relação ao nível de atividade econômica, mas, especificamente para produtos hortícolas, não foram encontradas referências.

3.4 ANÁLISE DE SÉRIES TEMPORAIS

Dentre os objetivos da análise de dados temporais estão entender a estrutura existente, identificar padrões e tendências e fazer previsões para o futuro.

Previsão por meio de modelos de séries temporais tem estudo consolidado em áreas como mercado financeiro e demanda de sistema de abastecimento de energia. Estudos têm sido realizados com o objetivo de encontrar um modelo que permita prever o preço futuro de ações com precisão, como Weng et al. (2022), que utilizou o método ARIMA para ações da Tesla. Um outro estudo, realizado por Vega-Márquez et al. (2021), utilizou o método LSTM para prever o consumo de energia, com bom desempenho, caracterizado pelo erro médio absoluto.

3.4.1 Métodos Clássicos

O método ARIMA (*Autoregressive Integrated Moving Average*) foi introduzido por Box e Jenkins em 1970 como uma técnica para modelar e prever séries temporais. Desde então, tornou-se um dos métodos mais populares e amplamente utilizados para análise em diferentes áreas, incluindo finanças, economia e agricultura (BOX et al., 2015).

Uma das principais vantagens do método ARIMA é sua capacidade de modelar séries temporais com diferentes padrões, como tendências, sazonalidades e ciclos. O método é capaz de lidar com dados que apresentam correlações entre os valores, o que é comum nos setores relacionados a agricultura (MAKRIDAKIS et al., 1998).

Esse método é baseado em três componentes principais: a parte autoregressiva (AR), a parte de médias móveis (MA) e a parte integrada (I). A parte AR é responsável por modelar a relação entre a série temporal atual e seus valores passados, enquanto a parte MA modela a relação entre a série temporal e seus erros passados. A parte I é usada para tornar a série estacionária, ou seja, remover tendências e sazonalidades (HYNDMAN; ATHANASOPOULOS, 2018).

O método autoregressivo assume que a série é estacionária, o que nem sempre ocorre. Além disso, valores extremos e mudanças abruptas podem prejudicar a capacidade preditiva do modelo (BOX et al., 2015; PÉREZ, 2021).

O modelo é definido por meio das Equações 1, 2 e 3, que correspondem as partes AR, MA e I, respectivamente.

$$AR(p): Xt = c + \Sigma \alpha_i X_{t-i} + \varepsilon_t \quad (1)$$

$$MA(d): X_t = c + \Sigma \beta_i \varepsilon_{t-i} + \varepsilon_t \quad (2)$$

$$I(q): X_t = (1 - B)^d (X_t - \mu) \quad (3)$$

Onde X_t é a série temporal; ε_t é o erro no período t; c é uma constante; B é o operador de deslocamento atrasado; μ é a média da série temporal; α_i e β_i são os coeficientes autoregressivos e de média móvel, respectivamente; e p, d e q são os parâmetros do modelo que indicam o número de termos autoregressivos, o número de diferenças e o número de termos de média móvel, respectivamente.

Os coeficientes α_i e β_i são estimados a partir dos dados, e a ordem p, d e q são determinadas por meio de técnicas como a função de autocorrelação e a função de autocorrelação parcial.

Para lidar com as limitações do método ARIMA quando há sazonalidade na série de dados, ou esses são multivariados, foram desenvolvidas variantes, como os métodos SARIMA (*Seasonal* ARIMA) e ARIMAX (*Moving Average with Explanatory variable*). Esse último inclui outras variáveis explicativas, além da própria série temporal (HYNDMAN; ATHANASOPOULOS, 2018).

O modelo SARIMA pode ser definido pela Equação 4.

$$SAR(s, P, D, Q): X_t = c + \Sigma \alpha_i X_{t-i} + \Sigma \varphi_j X_{t-s-j} + \Sigma \beta_i \varepsilon_{t-i} + \Sigma \theta_j \varepsilon_{t-s-j} + \varepsilon_t \quad (4)$$

Onde s é o período sazonal; o termo 'c' representa a constante ou o componente de deslocamento da série temporal; P, D e Q são os parâmetros do modelo que indicam o número de termos autoregressivos o número de diferenças sazonais e o número de termos de média móvel sazonais, respectivamente; e φ_j e θ_j são os coeficientes autoregressivos sazonais e de média móvel sazonais, respectivamente.

O modelo apresentado leva em consideração tanto os efeitos de curto prazo (AR, MA) quanto os efeitos de longo prazo (diferenciação) e sazonais (termos autoregressivos e de média móvel sazonais). Ao ajustar um modelo SARIMA aos dados, os coeficientes αi, φj, βi e θj são estimados usando métodos estatísticos e algoritmos de otimização, como o método dos mínimos quadrados ordinários (HYNDMAN; ATHANASOPOULOS, 2018).

O modelo ARIMAX pode ser definido pela Equação 5.

$$\text{ARIMA}(p, d, q): X_t = c + \Sigma\, \alpha_i X_{t-i} + \Sigma\, \beta_i\, \varepsilon_{t-i} + \beta_1 Z_{1t} + ... + \beta_k Z_{kt} + \varepsilon_t \quad (5)$$

Onde $\beta_1 Z_{1t} + ... + \beta_k Z_{kt}$ são os termos exógenos, representados por Z_{1t}, Z_{2t}, ..., Z_{kt}, onde k é o número de variáveis exógenas incluídas no modelo. Essas variáveis exógenas podem ser outras séries temporais ou fatores externos que se acredita afetarem a série temporal principal. Os coeficientes β_1, β_2, ..., β_k ponderam a contribuição dessas variáveis exógenas na previsão da série temporal.

Ao incluir variáveis exógenas no modelo ARIMA, espera-se que ele seja capaz de capturar melhor os padrões e as relações causais entre as variáveis, aumentando a precisão das previsões (BOX et al., 2015).

3.4.2 Regressão por Máquina de Vetor de Suporte

O método *Support Vector Regression* (SVR) é uma técnica de aprendizado de máquina que tem como objetivo encontrar a melhor função para descrever uma relação não linear entre as variáveis de entrada e saída de um conjunto de dados (SMOLA; SCHÖLKOPF, 2004; DRUCKER et al., 1997).

O SVR é uma extensão do algoritmo *Support Vector Machine* (SVM) utilizado para classificação, mas que foi adaptado para problemas de regressão (VAPNIK, 1995).

Para a tarefa de previsão é utilizado o conceito de máxima margem para encontrar a melhor linha de regressão para um conjunto de dados (VAPNIK, 1995). O método trabalha com a criação de um hiperplano em um espaço de alta dimensionalidade, de modo a maximizar a distância entre as amostras de dados mais próximas, conhecidas como vetores de suporte. O objetivo é encontrar o menor erro entre as previsões do modelo e os valores reais da variável de saída (SMOLA; SCHÖLKOPF, 2004).

Quanto a aplicação, pode ser utilizado para problemas de regressão univariados e multivariados, em diferentes áreas, como finanças, previsão de demanda, entre outras (XIE et al., 2006; MALDONADO et al., 2019).

Uma das principais vantagens do método SVR é sua capacidade de operar com dados não lineares e com alta dimensionalidade, o que pode ser um desafio para outros métodos de regressão. Além disso, ele é capaz de lidar com *outliers* e ruídos devido ao uso de funções de *kernel* (DRUCKER et al., 1997; SMOLA; SCHÖLKOPF, 2004).

O modelo pode ser representado pela Equação 6.

$$y = \sum (\alpha_i * K(x_i, x)) + b \quad (6)$$

Onde y é o valor previsto para a variável dependente; x é o vetor de entrada das variáveis independentes; α_i são os multiplicadores de Lagrange obtidos durante o treinamento do modelo; $K_{(xi,\ x)}$ é a função de kernel "rbf" (*Radial Basis Function*) que calcula a similaridade entre o vetor de características x_i de um exemplo de treinamento e o vetor de características x do exemplo de teste; e ‘b’ é o viés (*bias*).

A função de kernel "rbf" calcula a similaridade entre dois exemplos de dados com base em sua distância euclidiana. Quanto mais próximos os exemplos estiverem, maior será o valor da função (GÉRON, 2019).

Cada coeficiente de Lagrange α_i está associado a um exemplo de treinamento específico e determina o peso ou importância desse exemplo na construção da função de decisão. Esses coeficientes desempenham um papel crucial na determinação da posição e largura do hiperplano do SVR (MURPHY, 2012).

3.4.3 Memória de Curto e Longo Prazo

O *Long Short-Term Memory* (LSTM) é uma extensão das redes neurais recorrentes convencionais, que são uma classe de métodos de aprendizado de máquina projetados para lidar com dados sequenciais. O LSTM foi desenvolvido para resolver o problema de desvanecimento do gradiente, que ocorre quando os gradientes dos pesos da rede se tornam muito pequenos ou grandes durante o treinamento e dificultam a aprendizagem (HOCHREITER; SCHMIDHUBER, 1997).

O método usa um mecanismo de porta para controlar o fluxo de informação que entra e sai da célula de memória da rede. Essas portas são compostas de camadas de ativação, que

decidem qual informação deve ser esquecida ou mantida. Isso permite que o modelo mantenha dados relevantes por um longo período de tempo, enquanto descarta aqueles irrelevantes, melhorando assim a precisão (GERS et al., 2000).

As Equações 7 e 8 a seguir podem, de maneira geral, representar um modelo LSTM.

$$y_t = f(h_t) \tag{7}$$

$$h_t = LSTM(x_t, h_{\{t-1\}}) \tag{8}$$

Onde y_t é o valor de saída na posição t; f a função de saída; h_t o vetor de estado oculto na posição t; 'LSTM' é a camada da célula de memória; x_t o vetor de entrada na posição t; e $h_{\{t-1\}}$ é o vetor de estado oculto na posição t-1.

Na primeira equação, f é uma função de ativação aplicada ao estado oculto h_t para gerar a saída y_t. A função de ativação pode ser qualquer função adequada para o problema em questão, como uma função linear, sigmoide ou tangente hiperbólica.

A segunda equação atualiza o estado oculto da LSTM, o que ocorre em várias etapas, chamadas de "portões", que controlam o fluxo de informações dentro da célula da LSTM. Esses portões consistem em redes neurais com camadas de ativação específicas, que variam entre 0 e 1. O processo de atualização do estado oculto envolve a combinação de informações do estado anterior $h_{\{t-1\}}$ e da entrada atual x_t, que passam pelos portões para calcular o novo estado oculto h_t (GÉRON, 2019).

3.4.4 Redes Neurais Convolucionais

As redes neurais convolucionais (*Convolutional Neural Network* – CNN) são uma das técnicas mais poderosas para processamento de imagens e reconhecimento de padrões. Desde sua criação, em 1980, o método tem evoluído constantemente e revolucionado a área de visão computacional, proporcionando avanços significativos em tarefas como classificação de imagens, detecção de objetos, segmentação semântica, entre outras.

Alguns estudos foram realizados com o objetivo de explorar o potencial das CNNs para análise de séries temporais, com resultados promissores. De acordo com Lai et al. (2018), o método pode ser utilizado para modelar padrões temporais de curto e longo prazo.

O princípio básico das CNNs é a convolução, uma operação matemática que permite extrair características relevantes em diferentes níveis de abstração. A partir desse processo, as redes são capazes de identificar padrões.

A representação mais detalhada do modelo pode ser dada pelas Equações 9 e 10.

$$y = g(W_{d} \cdot z + b_{d}) \quad (9)$$

$$z = \text{flatten}(\text{pool}(a(\text{conv}(X; W_{c})))) \quad (10)$$

Onde conv(X; W_c) é a operação de convolução da matriz de entrada X com os filtros da camada convolucional W_c; g é a função de ativação na camada densa; a é a função de ativação aplicada na saída da convolução; pool é a operação de *pooling*; flatten é a operação de achatamento da saída da camada de *pooling*; W_d é a matriz de pesos da camada densa; b_d é o vetor de bias da camada densa; e z é a saída da camada densa.

A operação de convolução envolve a aplicação de um conjunto de filtros W_c, gerando uma representação. A camada de ativação introduz não-linearidades a esse resultado. Funções comuns de ativação incluem ReLU (*Rectified Linear Unit*), sigmoide e tangente hiperbólica (GOODFELLOW et al., 2016).

A operação de *pooling* é usada para reduzir a dimensionalidade da representação convolucional, preservando as características mais relevantes. A função *flatten* é usada para converter a representação em duas dimensões, geradas pela camada anterior, em um vetor unidimensional. Isso é necessário para conectar a camada de *pooling* à camada densa de regressão (GOODFELLOW et al., 2016).

3.4.5 Impulso Extremo de Gradiente

Chen e Guestrin (2016) apresentam o *eXtreme Gradient Boosting* (XGBoost**)** como um sistema escalável e eficiente para construção de árvores de decisão. Esse algoritmo de aprendizado de máquina tem se destacado em diversas aplicações, como classificação e regressão. O algoritmo utiliza um conjunto de árvores construídas sequencialmente, onde cada nova árvore busca corrigir os erros das árvores anteriores. Assim, o processo de construção das árvores é iterativo e visa minimizar a função de perda global, combinando as previsões.

A função do modelo pode ser representada matematicamente pela equação geral de regressão, apresentada na Equação 11.

$$Y = f(X) + \varepsilon \quad (11)$$

Onde Y é a variável dependente; X é o conjunto de variáveis independentes; f é a função do modelo; e ε é o erro aleatório. A função f é aproximada por uma combinação de árvores de decisão.

A função de perda utilizada pelo XGBoost é definida pela Equação 12.

$$L(y, f(x)) = \Sigma_i \, l(y_i, f_i(x_i)) + \Omega(f) \quad (12)$$

Onde L é a função de perda; y é a variável dependente; f(x) é a função do modelo; l é a função de perda por amostra; e Ω é a função de penalidade para complexidade do modelo.

'$l(y_i, f_i(x_i))$' é a função de perda que mede a discrepância entre o valor verdadeiro 'y_i' e a previsão do modelo '$f_i(x_i)$' para cada exemplo 'i' do conjunto de dados. Além da função de perda, o XGBoost também incorpora um termo de regularização, que é adicionado para evitar *overfitting*. Esse termo penaliza modelos mais complexos, favorecendo aqueles mais simples e, portanto, mais generalizáveis (CHEN; GUESTRIN, 2016).

3.4.6 Teste de Estacionariedade

O teste Dickey-Fuller Aumentado (ADF) é um dos métodos mais utilizados para testar a estacionariedade de uma série temporal. Ele é baseado na ideia de que uma série não estacionária pode ser transformada por meio de diferenciação. Esse teste foi desenvolvido por Dickey e Fuller (1979) e posteriormente aprimorado por Said e Dickey (1984).

O teste ADF é uma extensão do teste Dickey-Fuller original, que utiliza uma regressão linear para testar a presença de raízes unitárias em uma série temporal. Uma raiz unitária indica que a série não é estacionária e que há uma relação de longo prazo entre as observações. No teste, a regressão linear é aprimorada com o uso de um termo de correção que leva em conta a presença de autocorrelação nas observações (DICKEY; FULLER, 1979).

O resultado do teste é um valor estatístico denominado ‘estatística-t’, que é comparado a um conjunto de valores críticos para determinar se a série é estacionária ou não. O valor crítico é determinado com base no nível de confiança desejado e no tamanho da amostra (SAID; DICKEY, 1984).

3.4.7 Análise de Componentes Principais

A análise de componentes principais (ACP) é uma técnica estatística utilizada para reduzir a dimensionalidade dos dados, identificar padrões e relações entre as variáveis e extrair informações importantes de um conjunto de dados multivariados.

Com o aumento da quantidade de variáveis, a complexidade do modelo aumenta e a interpretação é dificultada. A ACP permite reduzir a dimensionalidade dos dados ao extrair um menor número de componentes que representam a maior parte da variação observada (WITTEN et al., 2016). Com a remoção de variáveis irrelevantes ou redundantes, a precisão do modelo melhora e a interpretação é facilitada (ZHANG et al., 2009).

Para Dillon e Goldstain (1984) essa análise ajuda a identificar padrões e relações entre as variáveis que não são facilmente observáveis com todos os fatores em conjunto. Assim, podem ser identificados grupos de variáveis que estejam altamente correlacionadas ou que contribuem significativamente para a variação nos dados.

A fórmula geral para obter o i-ésimo componente principal (CPi) é expressa pela Equação 13.

$$CPi = a_{i1} * X_1 + a_{i2} * X_2 + ... + a_{ip} * X_p \quad (13)$$

Onde X_j são as variáveis originais e a_{ij} são os coeficientes dos componentes principais, que são obtidos a partir dos autovetores da matriz de covariância ou de correlação das variáveis.

O primeiro componente principal é aquele que explica a maior parte da variância total dos dados, o segundo componente principal é aquele que explica a maior parte da variância restante, e assim por diante. O número de componentes principais é igual ou menor que o número de variáveis originais.

3.4.8 Análise de Resíduos

A análise de resíduos em séries temporais é uma ferramenta importante para verificar a adequação de um modelo aos dados observados. Os resíduos são as diferenças entre os valores reais da série e os valores ajustados pelo modelo. Se o modelo for adequado, os resíduos devem ser normais, homocedásticos e independentes (MORETTIN; TOLOI, 2018).

A normalidade refere-se à distribuição dos erros do modelo, que deve ser aproximadamente normal. A homocedasticidade significa que a variância dos erros é constante para todos os valores da variável independente. A independência implica que os erros não estão correlacionados entre si.

Segundo Baltar (2009), a normalidade e a homocedasticidade são desejáveis, pois facilitam a interpretação dos resultados e a realização de testes de hipóteses. No entanto, a independência é indispensável e eliminatória, pois se os erros forem dependentes, o modelo perde sua validade e confiabilidade.

Para testar essas propriedades podem ser utilizados testes estatísticos, como os testes de Ljung-Box, Jarque-Bera e o teste ARCH (*Autoregressive Conditional Heteroscedasticity*).

O teste de Ljung-Box é usado para testar a hipótese nula de que os resíduos não têm autocorrelação até uma certa ordem. O teste usa uma estatística baseada na soma dos quadrados das autocorrelações amostrais dos resíduos e segue uma distribuição qui-quadrado sob a hipótese nula. Se o valor-p do teste for menor que um nível de significância pré-definido, rejeita-se a hipótese nula e conclui-se que os resíduos têm autocorrelação (CRYER; CHAN, 2008).

A Equação 14 apresenta a estatística do teste.

$$Q = n * (n + 2) * \sum(r^2) / (n - k) \quad (14)$$

Onde Q é a estatística de teste; n é o número de observações da série temporal; k é o número de parâmetros estimados no modelo; e r^2 são as autocorrelações dos resíduos ao quadrado.

De acordo com Bera e Jarque (1980), os testes de Jarque-Bera e ARCH possuem algumas limitações ao serem aplicados a conjuntos de dados pequenos. O primeiro assume que os dados seguem uma distribuição normal e mostra-se mais adequado para conjuntos de dados com tamanhos maiores, onde as estimativas dos coeficientes de assimetria e curtose têm menor variabilidade. O teste ARCH é projetado para detectar a presença de heterocedasticidade condicional. O poder estatístico desse teste também diminui com tamanhos de amostra pequenos, tornando mais difícil a detecção confiável da condição analisada.

3.4.9 Análise de Importância

Segundo Lundberg e Lee, (2017), a análise de importância de variáveis é uma técnica utilizada em estudos de modelos de aprendizado de máquina para identificar quais variáveis preditoras têm maior influência sobre a variável resposta.

Os modelos de aprendizado de máquina são frequentemente considerados "caixas-pretas", pois suas decisões são baseadas em cálculos complexos e difíceis de interpretar. A análise de importância de variáveis pode tornar esses modelos mais explicáveis, fornecendo informações sobre como as variáveis contribuem para o resultado previsto.

Essa técnica também é útil para identificar problemas de sobreajuste e subajuste. Quando uma variável possui uma importância muito alta, pode indicar que o modelo está superajustado aos dados de treinamento, o que pode prejudicar sua capacidade de generalizar para novos dados. Por outro lado, se a importância for muito baixa, pode indicar subajuste, onde o modelo não está capturando adequadamente as relações presentes nos dados.

3.4.10 Teste Estatístico de Significância

O teste estatístico de significância desempenha um papel importante na comparação de métodos de previsão. Com isso, é possível obter uma medida objetiva para determinar se as diferenças observadas entre os métodos são estatisticamente significativas ou devidas ao acaso.

Uma das vantagens desse teste é que ele não exige uma distribuição específica dos dados, tornando-o adequado para análises não paramétricas. Além disso, ele pode ser aplicado a amostras de diferentes tamanhos e é robusto a valores discrepantes, conforme apresentado por Rosner (2006).

Rosner (2006) utilizou o teste de Wilcoxon, uma ferramenta estatística não paramétrica, para avaliar a diferença entre duas amostras independentes de dados emparelhados quando a suposição de normalidade dos dados não era atendida. O autor argumenta que essa é uma alternativa ao teste t de Student quando os dados não seguem uma distribuição normal ou quando a variável de interesse é de natureza ordinal.

Ao contrário do teste t, que compara as médias das duas amostras, o teste de Wilcoxon se concentra nas diferenças entre os pares de observações emparelhadas. As diferenças são classificadas em ordem e procede-se à avaliação se há significância estatística entre os grupos.

Para realizar o teste de Wilcoxon, os dados devem ser organizados em pares de observações. Em seguida, as diferenças entre essas observações são calculadas. As diferenças são classificadas em ordem absoluta e é calculado o valor da estatística de teste.

A fórmula geral do teste é dada pela Equação 15.

$$U = R - (n_1 * (n_1 + 1)) / 2 \quad (15)$$

Onde U é o valor estatístico do teste de Wilcoxon; R é a soma dos postos (ou ordens) da amostra menor; e n_1 é o tamanho da primeira amostra.

O valor de U é comparado a uma tabela de valores críticos para determinar a significância estatística do teste. A tabela de valores críticos depende do tamanho da amostra e do nível de significância desejado.

O resultado do teste de Wilcoxon é um valor de p, que representa a evidência estatística de diferença entre as amostras. Se o valor de p for menor que o nível de significância pré-determinado, como 0,05, pode-se rejeitar a hipótese nula e concluir que há diferença estatisticamente significativa entre as amostras (HIGGINS; GREEN, 2011).

3.5 TRABALHOS RELACIONADOS

Segundo Hamulczuk et al. (2021), para realizar da previsão de preços, é necessário considerar diversos fatores, como variáveis climáticas, a produção, o câmbio e as políticas governamentais. Os autores abordam o fenômeno da sazonalidade, definido como um movimento regular observado dentro do período analisado. Para eles a sazonalidade dos preços é uma consequência da variação da intensidade do trabalho, da oferta e do comércio no mercado. Assim, a flutuação da oferta, e consequentemente dos preços, tem sua origem no processo biológico de produção, que está estritamente relacionado à temperatura.

Nazlioglu e Soytas (2011) examinaram a interdependência de curto e longo prazo entre os preços mundiais do petróleo, a taxa de câmbio e os preços de commodities agrícolas na Turquia. Os resultados desse estudo sugerem que os preços agrícolas turcos não reagem significativamente aos choques do preço do petróleo e da taxa de câmbio no curto prazo. A análise de causalidade de longo prazo revela que as mudanças nos preços do petróleo e a valorização/depreciação da lira turca não são transmitidas aos preços das *commodities* agrícolas.

Kurumatani (2020) compara métodos baseados em LSTM, RNN e GRU para previsão de preços de tomate, repolho e alface no Japão. Os resultados mostraram que o LSTM apresentou o melhor desempenho em termos de precisão somente quando é treinado adequadamente por um número suficiente de épocas. Porém, com menos épocas de treinamento, o método RNN simples mostrou menor taxa de erro.

Paul et al. (2022) investigaram o uso de técnicas de aprendizado de máquina para prever o preço de atacado da berinjela em dezessete mercados principais em Odisha, Índia. Os autores compararam o desempenho dos métodos GRNN, SVR, RF e GBM com o ARIMA. Os resultados mostraram que os algoritmos de aprendizado de máquina superaram o modelo ARIMA em termos de precisão. O algoritmo GRNN foi o mais preciso, seguido pelo RF. Outra descoberta foi que a precisão dos algoritmos de aprendizado de máquina variava dependendo do mercado.

Purohit et al. (2021) aplicaram os métodos SVM, ANN, LSTM e ARIMA para prever o preço de cebola. Os resultados mostraram que, de acordo com as métricas RMSE e SMAPE, o método ARIMA foi o mais preciso.

Yoo e Oh (2020) discutem o uso de métodos baseados em LSTM, ARIMA, ETS e Prophet para prever os volumes de vendas de produtos agrícolas. O modelo resultante do primeiro método superou os outros em termos de precisão e foi capaz de capturar os padrões sazonais nos dados de vendas. Os autores também descobriram que a precisão do modelo LSTM variou entre os produtos.

O ARIMA foi aplicado por Darekar e Reddy (2017) na previsão do preço de mercado da soja nos principais estados indianos. Nesse estudo o modelo foi capaz de prever o preço com MAPE de 5,1%.

O método SARIMA foi a escolha de Reddy (2019) para prever o preço dos tomates na Índia. O autor coletou os preços mensais dos tomates de janeiro de 2006 a dezembro de 2016 para cinco estados do país. O modelo com menor erro percentual médio absoluto para as previsões, 18,1%, foi o SARIMA (2,0,0) (1,1,0).

Anggraeni et al. (2019) propõe a análise comparativa entre RNAs, ARIMA e ARIMAX para prever o preço do arroz na Indonésia. O último método utiliza variáveis exógenas, como a taxa de câmbio entre a rupia indonésia e o dólar americano, para melhorar a precisão das previsões. O modelo ARIMA obteve a melhor precisão, com MAPE) de 4,6%.

Shengwei et al. (2017) desenvolveram um método SVR aprimorado para prever preços de grãos na China. O método é baseado em um algoritmo de regressão que utiliza um conjunto de *kernels* para capturar a dinâmica do mercado. Foram considerados fatores

externos que afetam o preço, como a produção, condições climáticas, políticas governamentais, entre outros. Os autores usaram a análise de componentes principais (ACP) para reduzir a dimensionalidade das variáveis de entrada.

O estudo de Jin et al. (2019) utilizou o método LSTM para prever os preços de produtos hortículas na China. O modelo gerado superou os métodos clássicos em termos de precisão. Esse estudo demonstra a eficácia do método LSTM na manipulação de relacionamentos complexos e não lineares entre variáveis.

Da mesma forma, Ly et al. (2021) comparam o desempenho das LSTMs com o método de previsão tradicional ARIMA para prever os preços de algodão e óleo vegetal. Os autores coletaram dados sobre os preços junto ao Banco Mundial e realizaram previsões para um mês adiante. Os resultados indicam que os modelos de redes neurais não superam os tradicionais.

Santos et al. (2021) apresentam um modelo de previsão de preços do etanol *spot* brasileiro utilizando redes neurais artificiais com a arquitetura LSTM e o compara com previsões de SVM, SVR e RF. Os resultados alcançados mostram que a LSTM produziu os menores erros de regressão, no entanto, em relação à precisão da direção nas previsões, o SVM foi o melhor algoritmo para detectar tendências, obtendo bons resultados para todas as janelas de tempo utilizadas.

Cinco métodos baseados em aprendizado profundo, que usam diferentes variantes de LSTM, foram propostos por Murugesan et al. (2022) para prever os preços mensais de cinco *commodities* agrícolas. Os resultados mostraram que os modelos podem capturar a tendência e sazonalidade da série de preços e fornecer previsões razoáveis.

A literatura científica com referências sobre a aplicação do método CNN de regressão para previsão de preços ou produção agrícola ainda é restrita. Liu et al. (2022) propuseram um método de aprendizado profundo baseado em redes convolucionais para a previsão de preços de *commodities*. O modelo foi testado em dados de preços de soja, e os resultados mostraram um desempenho considerado satisfatório pelos autores, com captura de informações de curto e longo prazo de maneira eficiente.

Assim como para o método CNN, o XGBoost com aplicação em séries de preços agrícolas não dispõe de ampla base de estudos científicos. As abordagens mais recentes consideraram informações de mercado, clima, análise de sentimento e processamento de linguagem natural para obter uma maior precisão nas previsões.

O artigo de Yuan e Ling (2020) propõe um aplicativo de software com recursos de previsão de preços para ajudar os agricultores a obter conhecimentos sobre o mercado

agrícola e maximizar seus lucros. O artigo investiga diferentes algoritmos, como ARIMA, LSTM, SVR, Prophet e XGBoost. A partir do resultado, o LSTM foi o mais preciso e eficiente para lidar com grandes quantidades de dados complexos.

Guo et al. (2022) utilizaram modelos LSTM, RF, GBM e XGBoost para prever preços de milho. Concluíram que esses métodos apresentam alta precisão apenas quando os preços estão mudando levemente. Além disso, para o LSTM, foi identificado o problema da qualidade de histerese, quando o modelo mostra dificuldade em se ajustar rapidamente a mudanças bruscas nos preços.

Chen at al. (2022) propõem um novo método para a previsão de curto prazo de preços de tomate com uso de combinação de redes neurais convolucionais e modelos baseados em células LSTM. Eles argumentam que os preços agrícolas são influenciados por diversos fatores, como clima, e exibem padrões complexos que são difíceis de capturar por métodos clássicos. O modelo proposto foi comparado e superou os dois métodos isolados.

Devi et al. (2021) desenvolveram um método híbrido, que combina técnicas estatísticas e de aprendizado de máquina, incluindo média móvel integrada autorregressiva e redes neurais artificiais, para prever a produção de trigo, o que resultou estimativas precisas e de alta qualidade.

4 MATERIAL E MÉTODOS

O presente estudo adotou o Google Colaboratory, uma plataforma colaborativa baseada em nuvem, utilizada na área de ciência de dados e aprendizado de máquina. O uso dessa plataforma é motivada pela facilidade de acesso, flexibilidade e possibilidade de utilização de bibliotecas de código aberto para análise de dados. Com essa ferramenta, foi possível realizar a integração, análise exploratória, bem como a modelagem de dados.

4.1 CARACTERIZAÇÃO DOS DADOS E PRÉ-PROCESSAMENTO

Neste estudo foram utilizadas as séries temporais disponíveis para as cidades de São Paulo e Curitiba. Quando não houve dados disponíveis para os municípios, foram empregadas séries a nível dos Estados e Brasil.

As cidades selecionadas são as maiores capitais das regiões Sudeste e Sul do Brasil e representam mercados importantes para o consumo e a produção de tomate. Dados históricos confiáveis e atualizados para as séries de interesse estão disponíveis. As diferenças climáticas, logísticas e socioeconômicas que podem influenciar na formação e na dinâmica dos preços do tomate são interessantes para a discussão objeto deste trabalho.

Oito bases de dados foram utilizadas, compreendidas entre janeiro do ano de 2010 e dezembro do ano de 2021, obtidas em diversas fontes, como apresentado na Tabela 2.

Tabela 2 - Descrição, fontes e número de instâncias das bases de dados utilizadas no estudo

Descrição do dado	**Tipo de acesso**	**Fonte**	**N.° de instâncias**
Médias mensais, no atacado, de preço e quantidade comercializada de tomate	Público	Conab	4095795
Dados meteorológicos	Público	Inmet	91980
Dados sobre atividade do comércio amplo	Público	IBGE	2340
Preços de combustíveis	Público	ANP	23328
Média do valor do dólar americano	Público	Banco Central	144
Taxa do Índice de Preços ao Consumidor Amplo (IPCA)	Público	IBGE	2160
Média de preços de fertilizantes	Público	Conab	1584
Médias do preço do tomate pago aos agricultores	Público	Conab	3744
		Deral-PR	144

Os dados das médias mensais de preço no atacado e quantidade comercializada de tomate no Ceagesp e Ceasa/PR foram obtidos após ser aplicada filtragem na base de dados primários. Todas as categorias de tomate foram consideradas, sem seleção por classe e tipo.

O segundo conjunto de dados inclui informações sobre dados meteorológicos registrados a cada hora pelas estações automáticas do Inmet, localizadas nos municípios de São Paulo e Curitiba, uma em cada localização. A dimensão das bases foi reduzida, com manutenção dos registros de temperatura e precipitação. Essas duas variáveis climáticas foram selecionadas como possíveis interferentes no comportamento alimentar da população dos locais analisados.

Os dados mensais sobre a atividade do comércio nos dois locais são públicos e estão disponíveis no sítio do Instituto Brasileiro de Geografia e Estatística (IBGE) na internet.

Os preços mensais de óleo diesel nas cidades de São Paulo e Curitiba também são públicos e estão disponíveis no sítio da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) na internet. A base possui informações para diferentes tipos de combustíveis, por isso, foi realiza a filtragem.

Os preços médios, em Reais, a cada mês, do dólar no Brasil estão disponíveis no sítio do Banco Central do Brasil na internet.

Os dados mensais sobre o Índice de Preços ao Consumidor Amplo (IPCA) no Brasil estão disponíveis no sítio do IBGE na internet. Eles incluem informações sobre a variação de preços dos principais serviços e produtos que compõem o custo de vida do brasileiro médio.

A fonte de dados de preços do fertilizante ureia, um dos insumos mais utilizados na agricultura, foi a Conab. A Companhia realiza regularmente pesquisas de preços em diversas regiões do país. A base contém registros para diferentes fertilizantes e Estados do Brasil, por isso, foi aplicada a filtragem.

A CONAB também é a fonte dos dados públicos para a variável preços pagos aos produtores de tomate ao longo do tempo no Estado de São Paulo. Essa base, por conter informações sobre diferentes Estados, foi filtrada para os registros de interesse. Os preços praticados no Paraná foram obtidos no sítio do Deral/PR (Departamento de Economia Rural do Estado do Paraná) na internet.

Os dados brutos, geralmente, contêm erros, informações faltantes e outras inconsistências que podem afetar negativamente a qualidade dos resultados da análise (MCKINNEY, 2017). A limpeza primária é uma etapa crítica no processo de análise, que envolve a detecção e correção de problemas nos registros coletados.

A não correção das séries temporais de dados econômicos pelo índice de inflação do período foi uma escolha deliberada. Esse ajuste teria o efeito de remover a variação nominal, tornando os dados ajustados em termos reais. No entanto, para a aplicação dos métodos, é interessante considerar a variação nominal, uma vez que os modelos devem ser capazes de lidar com as flutuações reais.

Para integrar, manipular e realizar o pré-processamento dos dados, foi utilizada a biblioteca Pandas, da linguagem de programação Python. Essa biblioteca de análise de dados fornece uma estrutura de fácil utilização e flexível para trabalhar com tabelas de dados, permitindo que o usuário realize várias operações, como seleção de dados, filtragem, agregação, transformação e limpeza (PANDAS, 2023).

Uma das principais ferramentas oferecidas pelo Pandas é o objeto *DataFrame*, uma estrutura que armazena dados em uma tabela bidimensional. Essa tabela pode ser facilmente manipulada para realizar operações como a remoção de valores duplicados, a eliminação de linhas ou colunas com valores nulos, correção de valores inválidos e conversão de tipos de dados (VANDERPLAS, 2016).

Para realizar a limpeza primária dos dados, foram adotados os seguintes procedimentos:

- Verificação de dados duplicados: verificação de linhas com informações idênticas, para evitar que dados redundantes comprometam a análise. Os registros duplicados foram eliminados;
- Verificação de valores nulos: o tratamento de registros com informações faltantes é importante para garantir a precisão da análise. Os registros que apresentaram valores nulos foram substituídos por médias.
- Verificação de valores extremos: a verificação de valores extremos é importante para evitar que dados discrepantes comprometam a análise. Para isso, foram calculados os valores mínimos e máximos para cada variável. Caso necessário, o valor foi substituído pela média.
- Verificação de valores inconsistentes: a verificação de valores inconsistentes garante a integridade dos dados. Os registros que apresentaram informações que não fazem sentido ou que não condizem com a realidade, foram substituídos.

Para todas as bases com frequência diferente de valores mensais, foi aplicada a técnica de agregação.

O método *seasonal_decompose* da biblioteca *Statsmodels.tsa.seasonal* foi utilizado para decompor as séries de preço de tomate em suas componentes sazonais, tendência e resíduos.

A ferramenta *plot_acf* da biblioteca *statsmodels.graphics.tsaplot*s foi utilizada para plotar a função de autocorrelação, uma medida da relação entre observações separadas por um determinado intervalo de tempo. Com esse gráfico é possível identificar padrões de correlação entre observações em diferentes atrasos, o que pode auxiliar no entendimento da dinâmica da série temporal e a discutir modelos utilizados para a previsão (BOX et al., 2015).

4.2 NORMALIZAÇÃO DOS DADOS

A normalização de dados é uma técnica utilizada em diversas áreas, incluindo a análise de dados e o aprendizado de máquina. Essa etapa é importante para garantir que as variáveis em um conjunto de dados tenham a mesma escala, o que pode melhorar a performance dos modelos de previsão (SCIKIT-LEARN, 2023). Neste estudo foi utilizado o método de normalização de dados da classe *MinMaxScaler()*, do pacote Scikit-learn, da linguagem de programação python.

O *MinMaxScaler()* é uma técnica de normalização que redimensiona as variáveis de um conjunto de dados para um intervalo de 0 a 1. Isso é feito através da Equação 16.

$$x' = (x - \min(x)) / (\max(x) - \min(x)) \quad (16)$$

Onde x é a variável original, x’ é a variável normalizada, min(x) é o menor valor da variável original e max(x) seu maior valor.

4.3 TESTE ADF

A estacionariedade facilita a modelagem e a interpretação dos dados, pois permite assumir que o padrão observado no passado se mantém no futuro. Para avaliar a estacionariedade das séries utilizadas neste estudo, foi utilizado o teste Dickey-Fuller Aumentado (ADF), disponível no pacote *statsmodels.tsa.stattools*, da linguagem de programação python.

O código utilizado para realizar o teste consistiu em uma função que recebe como entrada a série temporal a ser analisada. Dentro da função, foram determinadas as estatísticas

de média móvel e desvio padrão móvel, com uma janela de 12 observações, dada a previsão de sazonalidade anual.

O teste Dickey-Fuller foi então realizado, e os resultados apresentados em forma de tabela, contendo o valor do teste estatístico, o valor-p, o número de atrasos e observações utilizadas. Além disso, foram fornecidos os valores críticos para diferentes níveis de significância.

Caso a estatística do teste indique que a série é não-estacionária, foi aplicada a técnica de diferenciação para torná-la estacionária antes de gerar os modelos. Essa técnica é uma abordagem comum para remover tendências e tornar uma série temporal estacionária (MORETTIN; TOLOI, 2018).

A diferenciação dos dados envolve a subtração dos valores da série temporal em um determinado período de tempo pelo valor no período anterior. A primeira diferença é a aquela entre o valor da série temporal no tempo t e o valor da série temporal no tempo t-1. A segunda diferença é a que ocorre entre a primeira diferença no tempo t e a primeira diferença no tempo t-1, e assim por diante (BOX et al., 2015).

4.4 MÉTODOS PARA PREVISÃO

Para todos os métodos aplicados neste estudo, os dados coletados entre janeiro de 2010 e junho de 2021 foram utilizados como dados de treinamento. Aqueles compreendidos entre julho e dezembro de 2021, aplicados aos testes.

Ao utilizar dados mais recentes para avaliar a precisão dos modelos, pode-se ter uma visão mais atualizada e precisa do desempenho dos mesmos, com *insights* sobre a evolução recente dos preços, o que pode ser especialmente importante para tomadores de decisão em tempo real, como agricultores e comerciantes de tomate. Além disso, escolher um período de teste mais recente também é relevante para avaliar a capacidade do modelo de lidar com variações sazonais e flutuações econômicas recentes, que podem afetar os preços.

4.4.1 Aplicação dos Métodos Clássicos

Para a aplicação dos métodos clássicos (ARIMA, SARIMA e ARIMAX) foi utilizada a linguagem Python, com a classe ARIMA da biblioteca *statsmodels*, especificando a ordem das componentes ARIMA e as variáveis exógenas, quando for o caso, utilizadas no modelo.

Para definição dos parâmetros do modelo foi utilizada a função de ajuste automático chamada *auto_arima*. Essa função usa uma abordagem de busca em grade para testar diferentes combinações de parâmetros e seleciona o melhor modelo com base na medida de desempenho AIC (Critério de Informação de Akaike).

4.4.2 Aplicação do *Support Vector Regression*

Foi implementado um método SVR utilizando a linguagem Python. A classe SVR da biblioteca *scikit-learn*, especificando o *kernel* utilizado, os parâmetros de regularização e outros hiperparâmetros relevantes. A busca de grade (*grid search*) foi aplicada para encontrar os melhores hiperparâmetros candidatos para o modelo.

4.4.3 Aplicação do *Long Short-Term Memory*

O método LSTM foi aplicado em Python com uso da biblioteca *Keras*, onde é criado um modelo com a classe *Sequential* e adicionadas camadas LSTM, além de outras camadas, como densas e de *dropout*. Os hiperparâmetros foram ajustados a partir daqueles sugeridos pela função *GridSearchCV*, do *Scikit-Learn*.

4.4.4 Aplicação do *Convolutional Neural Network*

O método de rede neural convolucional de regressão foi aplicado com o pacote *tensorflow.keras.models*, da linguagem de programação Python. A construção do modelo envolve a definição da arquitetura da rede e a compilação. Os hiperparâmetros foram ajustados a partir da função *GridSearchCV*, do *Scikit-Learn*.

A definição da arquitetura da rede começa com a criação de um objeto *Sequential*, que permite empilhar camadas. Em seguida, são adicionadas as camadas de convolução. Essa foi definida com a função *Conv1D*, que recebe como parâmetros o número de filtros, o tamanho do *kernel*, a função de ativação e outras configurações.

O otimizador foi definido com a função *Adam*, que é uma versão aprimorada do algoritmo de descida do gradiente estocástico.

4.4.5 Aplicação do *eXtreme Gradient Boosting*

O XGBoost foi implementado com o uso da biblioteca *XGBoost*, do pacote *xgb.XGBRegressor*, da linguagem de programação Python. Essa biblioteca oferece várias opções para ajustar os hiperparâmetros, como a profundidade da árvore, a taxa de aprendizado e a regularização. Os hiperparâmetros foram ajustados a partir da função *GridSearchCV*, do *Scikit-Learn*.

4.5 MÉTRICAS DE DESEMPENHO

As métricas de desempenho são essenciais na avaliação da qualidade de um modelo de previsão ou classificação. Entre as mais utilizadas estão a raiz quadrada do erro médio (*Root Mean Squared Error* - RMSE) e a porcentagem absoluta do erro médio (*Mean Absolute Percentage Error* – MAPE), que foram utilizadas neste estudo.

O RMSE é uma medida de dispersão que indica o quão distantes os valores previstos estão dos valores reais. Ele é calculado pela raiz quadrada da média dos erros ao quadrado. Uma vantagem do RMSE é que ele penaliza erros grandes.

O RMSE pode ser expresso pela Equação 17.

$$RMSE = sqrt(1/n * sum((y\_real - y\_pred)^2)) \quad (17)$$

Onde y_real é o valor observado, y_pred é o valor previsto e n é o número total de observações.

O MAPE é uma medida de precisão que indica o erro percentual médio entre os valores previstos e os valores reais. Ele é calculado pela média da diferença absoluta entre esses números, dividida pelo valor real. O MAPE pode ser expresso pela Equação 18.

$$MAPE = med(abs((y\_real - y\_pred) / y\_real)) \quad (18)$$

Onde y_real é o valor real, y_pred é o valor previsto e ‘med’ é a média dos valores.

A interpretação dos intervalos do MAPE neste estudo será aquela indicada por Makridakis et al. (1998). Valores menores ou iguais a 10% indicam que o modelo tem uma boa precisão na previsão dos valores futuros. Resultados maiores do que 10% e menores que

20% foram considerados razoáveis. Quando a métrica apontar valores acima de 20%, o modelo será classificado como de baixa precisão.

RMSE e MAPE foram calculadas em Python com o uso das bibliotecas NumPy, que é uma biblioteca para processamento numérico em Python.

Como os métodos LSTM e CNN são estocásticos, utilizam pesos aleatórios para inicializar os parâmetros da rede e atualizam esses pesos com base em um algoritmo de otimização estocástica, as simulações foram repetidas cinquenta vezes para cada conjunto de dados e método para medir a precisão, como sugerido por Purohit et al. (2021).

4.6 ANÁLISE DE COMPONENTES PRINCIPAIS

A ACP é útil para analisar dados com variáveis correlacionadas, pois permite identificar as dimensões latentes que estruturam os dados, simplificar a visualização em um espaço de menor dimensão e reduzir o ruído e a multicolinearidade.

Para a análise de componentes principais foram utilizadas as bibliotecas *NumPy* e *Scikit-learn*. Além disso, foi utilizada a biblioteca *Matplotlib* para a plotagem de gráficos.

Neste estudo foi utilizado o critério proposto por Johnson e Wichern (2014) para escolha do valor mínimo de variância explicada para determinar o número de componentes principais. Eles argumentam que um critério razoável para a escolha do número de fatores é reter aqueles que juntos expliquem pelo menos 80% da variância total.

4.7 DEFINIÇÃO DE HIPERPARÂMETROS

Neste estudo foram adotadas as configurações dos parâmetros mostradas na Tabela 3. Para os métodos SVR, LSTM, CNN e XGBoost foi aplicado atraso de 3 (*lag* = 3) períodos para as variáveis preditoras.

Tabela 3 - Hiperparâmetros para os modelos candidatos

(continua)

Modelo	**Hiperparâmetro**	**São Paulo**	**Curitiba**
ARIMA	*p*	3	2
	d	0	0
	q	0	1
SARIMA	*p*	3	2
	d	0	0

Tabela 3 - Hiperparâmetros para os modelos candidatos

(conclusão)

Modelo	Hiperparâmetro	São Paulo	Curitiba
SARIMA	*q*	0	0
	P	0	1
	D	1	3
	Q	1	0
	seasonal_order	12	12
ARIMAX	*P*	3	2
	D	0	0
	Q	0	1
SVR	*kernel*	'rbf’	'rbf’
	C	4,5	0,7
	gamma	0,5	0,9
LSTM	*units*	64	64
	activation	'relu’	'relu’
	batch_size	16	32
	epoch	100	200
	optimizer	'adam’	'adam’
CNN	*units*	64	64
	activation	'relu’	'relu’
	batch_size	32	16
	epoch	100	100
	optimizer	'adam’	'adam’
	pool_size	1	1
	verbose	0.02	1
XGBoost	*colsample_bytree*	1	1
	learning_rate	0.02	0,03
	max_depth	3	3
	objective	“reg:squarederror”	“reg:squarederror”
	reg_lambda	0,2	2,5
	subsample	0,4	0,2
	n_estimators	150	150

Fonte: O autor

4.8 TESTE DE LJUNG-BOX

No contexto deste estudo a atenção foi direcionada à análise de autocorrelação dos resíduos usando o teste de Ljung-Box. A limitação do tamanho da amostra pode levar a resultados imprecisos nos testes de normalidade e heterocedasticidade. Portanto, optou-se por não realizar esses testes específicos, a fim de evitar conclusões incorretas.

O teste de Ljung-Box foi realizado utilizando a função *acorr_ljungbox* da biblioteca estatística da linguagem de programação Python.

4.9 CÁLCULO DE IMPORTÂNCIA DAS VARIÁVEIS

Para o cálculo de importância das variáveis foi utilizada a ferramenta *permutation_importance*, disponível na biblioteca *sklearn, da* linguagem de programação Python. Com isso, foi medida a importância de cada variável por meio do cálculo da perda de acurácia do modelo. Essa função recebe como argumentos o modelo de aprendizado de máquina treinado, a matriz de dados com as variáveis preditoras e o vetor de dados com a variável resposta. Além disso, é possível definir o número de vezes que cada variável será embaralhada aleatoriamente através do parâmetro *n_repetitions*, e também especificar uma semente (*random_state*) para a geração de números aleatórios (SCIKIT-LEARN, 2021).

Os resultados obtidos com a função *permutation importance* indicam a importância relativa de cada variável em relação ao desempenho do modelo. Valores positivos significam que a permutação aleatória da variável resultou em uma queda no desempenho do modelo. Isso sugere que a variável original está fornecendo informações úteis, e seu embaralhamento leva a uma perda de precisão. De maneira contraria são interpretados os valores negativos.

Para o método XGBoost foi utilizado o atributo *feature_importances_* do modelo resultante. Quanto maior o valor de importância de uma variável, mais ela contribui para o desempenho do modelo.

4.10 TESTE DE WILCOXON

O teste de Wilcoxon é um método estatístico não paramétrico utilizado para comparar duas amostras relacionadas. É uma alternativa robusta quando os dados não atendem aos pressupostos dos testes paramétricos, como a distribuição normal dos dados ou a igualdade de variâncias. O teste de Wilcoxon avalia se existe uma diferença significativa entre os valores pareados de duas amostras (KAUR; MISHRA, 2018).

Neste estudo, foi utilizado o módulo *scipy*, uma biblioteca científica adotada em Python, para realizar o teste de Wilcoxon aplicado aos erros absolutos dos modelos analisados. O módulo oferece uma função chamada *wilcoxon*, que implementa o teste e facilita sua aplicação.

5 RESULTADOS E DISCUSSÃO

Após o pré-processamento dos dados, foram produzidas as dez séries temporais consideradas para este estudo: quantidades mensais de tomate comercializado, volumes mensais de precipitação, médias mensais de temperatura, taxas mensais do índice de preços ao consumidor, valores mensais para atividade do comércio, médias mensais de preço do Dólar, médias mensais de preço do fertilizante ureia, médias mensais de preços pagos aos produtores de tomate, médias mensais de preços praticados para o óleo diesel, médias mensais de preços no atacado do tomate no Ceagesp de São Paulo e no Ceasa/PR de Curitiba.

Inicialmente, foram realizadas análises descritivas dos dados, com a obtenção de medidas de tendência central e de dispersão.

As Tabelas 4 e 5 apresentam estatísticas descritivas (média; desvio padrão; valores mínimo e máximo; e percentis 25%, 50% e 75%) para as séries temporais analisadas neste estudo.

Tabela 4 - Estatística descritiva das variáveis consideradas para São Paulo

Estatística	Qtd	Pct	Temp	IPCA	Atv	Dolar	Ureia	Prc_pr	Diesel	Prec_to
Média	22639700,00	129,34	19,96	0,49	61,20	3,15	1606,74	2,01	2,84	3,00
Desvio padrão	5040152,00	105,69	2,33	0,34	15,53	1,20	602,63	0,77	0,75	1,03
valor mínimo	8372016,00	0,20	14,84	-0,38	32,66	1,57	853,00	0,66	1,96	1,35
25%	20439110,00	45,80	17,95	0,25	50,55	2,04	1313,75	1,47	2,10	2,20
50%	23312840,00	106,10	20,11	0,46	60,80	3,17	1480,00	1,85	2,90	2,89
75%	25888450,00	182,45	21,90	0,75	68,76	3,87	1770,81	2,41	3,31	3,69
Valor máximo	33218100,00	493,80	24,72	1,35	114,49	5,82	4879,98	4,30	5,28	6,88

Fonte: O autor

Tabela 5 - Estatística descritiva das variáveis consideradas para Curitiba

(continua)

Estatística	Qtd	Pct	Temp	IPCA	Atv	Dolar	Ureia	Prc_pr	Diesel	Prec_to
Média	4934409,00	121,49	17,71	0,49	61,96	3,15	1444,01	2,05	2,78	2,24
Desvio padrão	1323584,00	80,72	2,52	0,34	15,86	1,20	672,72	0,80	0,70	0,79
valor mínimo	454419,00	4,20	12,28	-0,38	31,22	1,57	781,47	0,66	1,94	0,66

Tabela 5 - Estatística descritiva das variáveis consideradas para Curitiba

(conclusão)

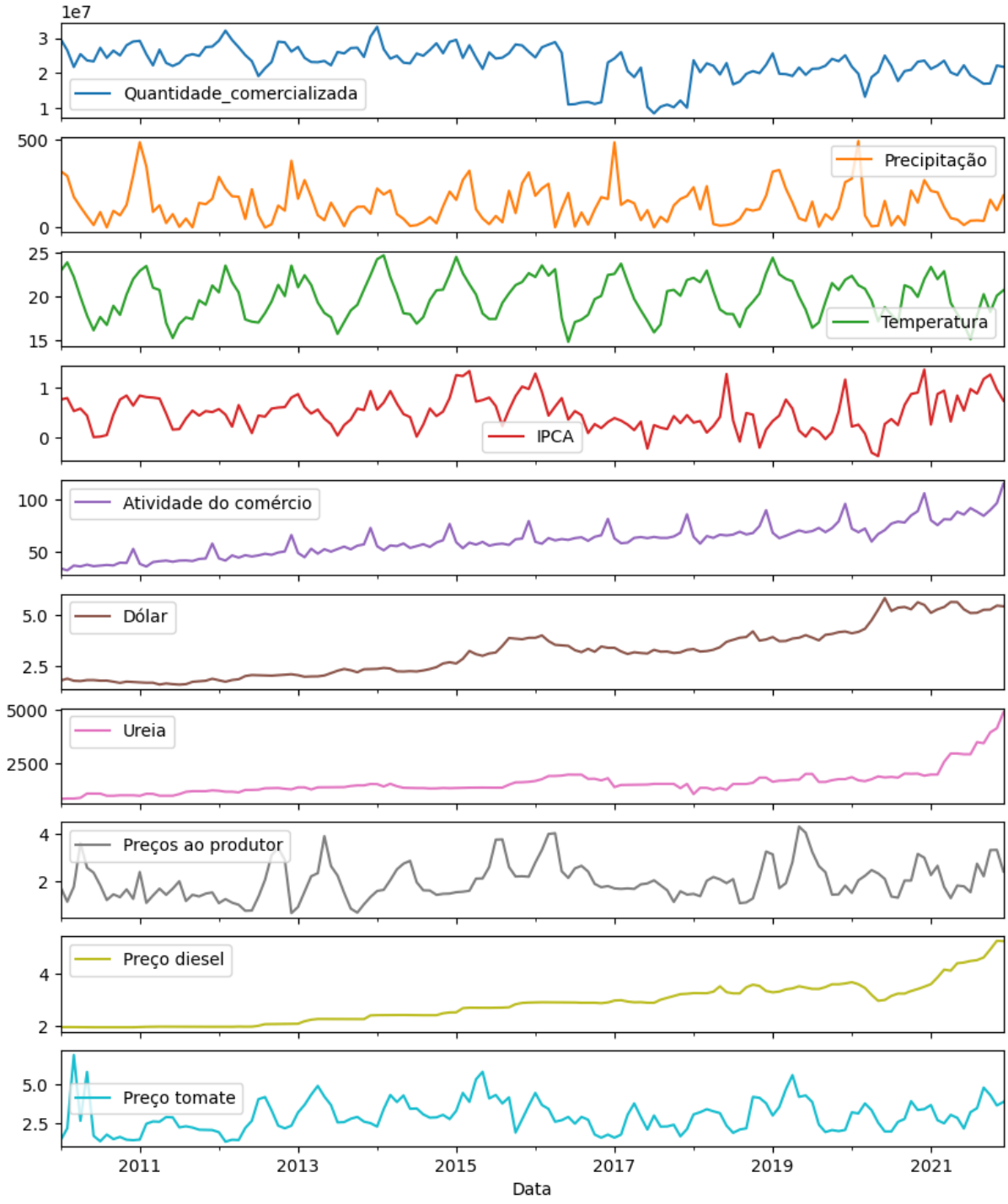
Estatística	Qtd	Pct	Temp	IPCA	Atv	Dolar	Ureia	Prc_pr	Diesel	Prec_to
25%	4499986,00	60,15	15,55	0,25	50,75	2,04	1234,44	1,46	2,09	1,64
50%	5215436,00	102,20	17,77	0,46	63,37	3,17	1301,11	2,00	2,78	2,38
75%	5778222,00	173,95	19,89	0,75	70,90	3,87	1397,66	2,61	3,21	2,72
Valor máximo	8493108,00	366,60	23,03	1,35	112,74	5,82	5233,48	5,06	5,11	4,42

Fonte: O autor

Nota: Para as Tabelas 4 e 5 considerar as abreviaturas: quantidades mensais de tomate comercializado (Qtd), volumes mensais de precipitação (Pct), médias mensais de temperatura (Temp), taxas mensais do Índice de Preços ao Consumidor (IPCA), valores mensais para atividade do comércio (Atv), médias mensais de preço do Dólar (Dolar), médias mensais de preço do fertilizante ureia (Ureia), médias mensais de preços pagos aos produtores de tomate (Pcr_pr), médias mensais de preços praticados para o óleo diesel (Diesel), médias mensais de preços do tomate nos Ceasas (Prec_to).

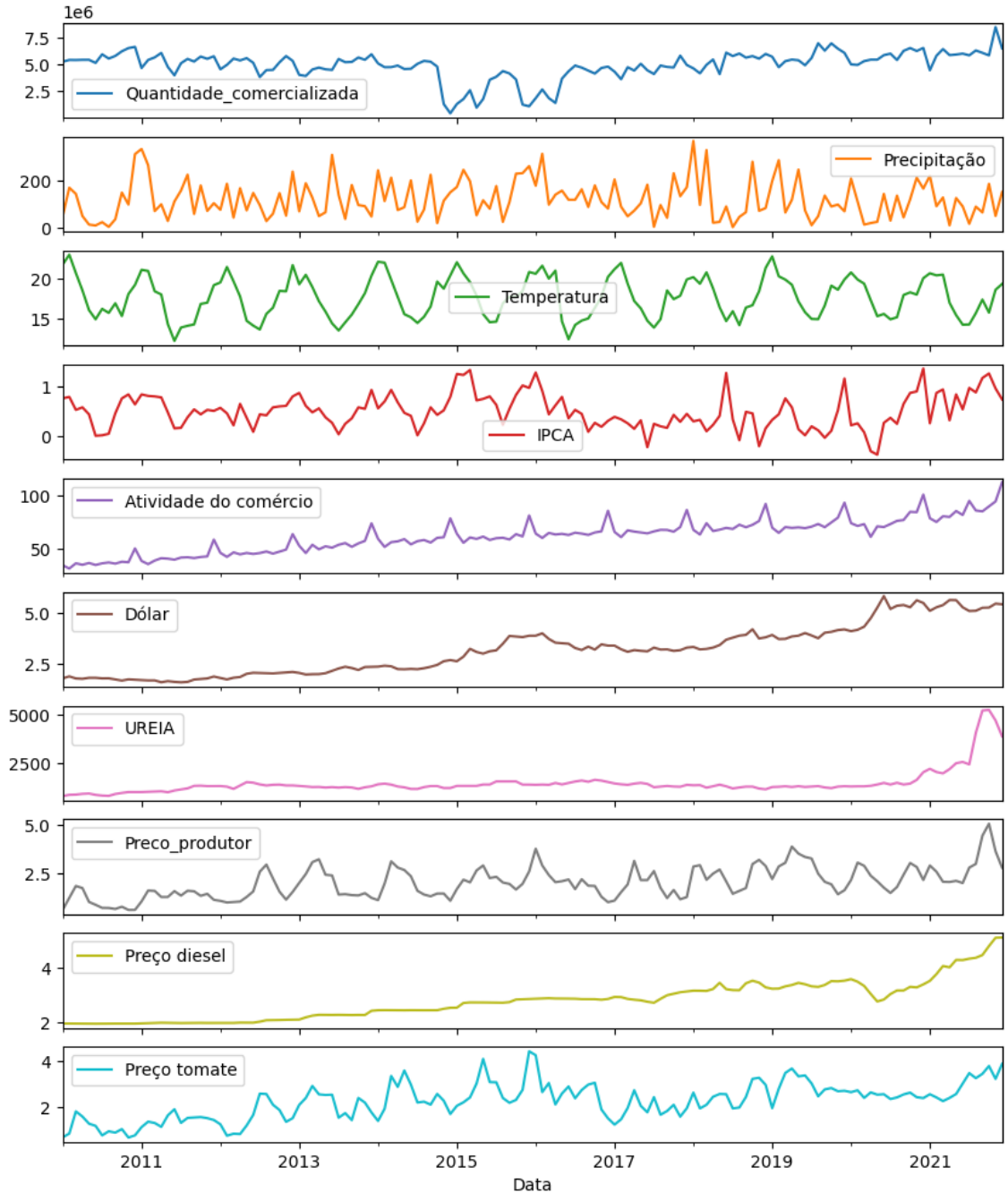
Para visualizar os dados das séries temporais aplicadas, foram plotados os Gráficos 1 e 2, que mostram a evolução dos registros ao longo do tempo nos dois locais estudados.

Gráfico 1 - Séries utilizadas no estudo para São Paulo. No eixo 'x' é apresentado o período de tempo em anos. No eixo 'y' estão plotados os valores não normalizados das variáveis



Fonte: O autor

Gráfico 2 - Séries utilizadas no estudo para Curitiba. No eixo 'x' é apresentado o período de tempo em anos. No eixo 'y' estão plotados os valores não normalizados das variáveis



Fonte: O autor

Em São Paulo, a estação chuvosa ocorre entre os meses de outubro e março, enquanto a estação seca vai de abril a setembro. Em Curitiba, a distribuição de chuvas é mais homogênea ao longo do ano, mas os meses mais chuvosos são de dezembro a fevereiro.

Outro ponto a se destacar é a grande variação na quantidade de chuva entre os anos, em ambas as cidades. Por exemplo, em São Paulo, a precipitação em janeiro de 2010 foi de 322,6 mm, enquanto em janeiro de 2013 foi de apenas 164,8 mm. Em Curitiba, a precipitação em julho de 2013 foi de 139,8 mm, enquanto em julho de 2014 foi de apenas 27,2 mm. A

precipitação também apresentou variação ao longo dos anos. Em São Paulo, os anos de 2010, 2011 e 2015 apresentaram uma quantidade maior de precipitação, enquanto em Curitiba, os anos de 2011, 2015 e 2020 tiveram a maior quantidade.

É possível visualizar que as temperaturas em São Paulo são, em média, mais altas do que em Curitiba, e, também, apresentam uma variação menor ao longo do tempo, indicado pelo desvio padrão.

Em relação às tendências ao longo do período analisado para o índice de atividade do comércio, houve um aumento gradual na atividade comercial em ambas as cidades até 2014, seguido por um período de queda até 2016 e uma retomada no crescimento até 2019. Em 2020, devido à pandemia de COVID-19, houve uma queda acentuada em ambos os locais, com recuperação gradual em 2021. É interessante notar que a atividade comercial em Curitiba e São Paulo parecem estar fortemente correlacionadas, seguindo padrões semelhantes ao longo do período. Há um padrão sazonal, com um pico no final do ano (novembro e dezembro) e queda no início do ano (janeiro e fevereiro). Isso pode estar relacionado a fatores sazonais, como o aumento de vendas de fim de ano e o período de férias no início do ano.

A série de preço do dólar é volátil, com altos e baixos frequentes ao longo dos anos. Em geral, a tendência parece ser de aumento, o que significa que o real perdeu valor em relação ao dólar ao longo do tempo. Houve um aumento constante na taxa de câmbio até 2014. Em 2015, a taxa subiu rapidamente para mais de 3 Reais por dólar. Nos anos seguintes, a taxa de câmbio continuou volátil, com um aumento significativo em 2020, provavelmente, como reflexo da pandemia de COVID-19, que causou queda na economia global e um aumento na demanda por dólares americanos como ativo (ANAYA, 2020).

Os preços do diesel aumentaram consistentemente ao longo dos anos, embora tenha havido flutuações mensais. Houve um aumento significativo nos preços a partir de 2018, com uma pequena queda em 2020. No entanto, a tendência geral é de alta.

As séries de preços do fertilizante ureia apresentam uma tendência de crescimento ao longo do tempo, com algumas oscilações sazonais, com pico no mês de dezembro e uma queda no mês de maio. Essa sazonalidade pode ser explicada pelo ciclo de produção agrícola, que tem seu pico no final do ano e uma queda na época de plantio.

Podemos observar que os preços pagos aos produtores de tomate em São Paulo e Curitiba parecem seguir um padrão semelhante ao longo do tempo. Ambos tiveram uma tendência de alta no início do período (2010-2013), seguida por uma queda acentuada em 2014, uma recuperação em 2015 e uma queda novamente em 2016. A partir de 2017, os preços tiveram uma tendência de alta até atingir um pico em 2021. O padrão sazonal também

é semelhante. Ambos apresentam um aumento nos preços durante os meses de junho a agosto e uma queda durante os meses de novembro a janeiro. Isso pode estar relacionado ao ciclo de produção e colheita.

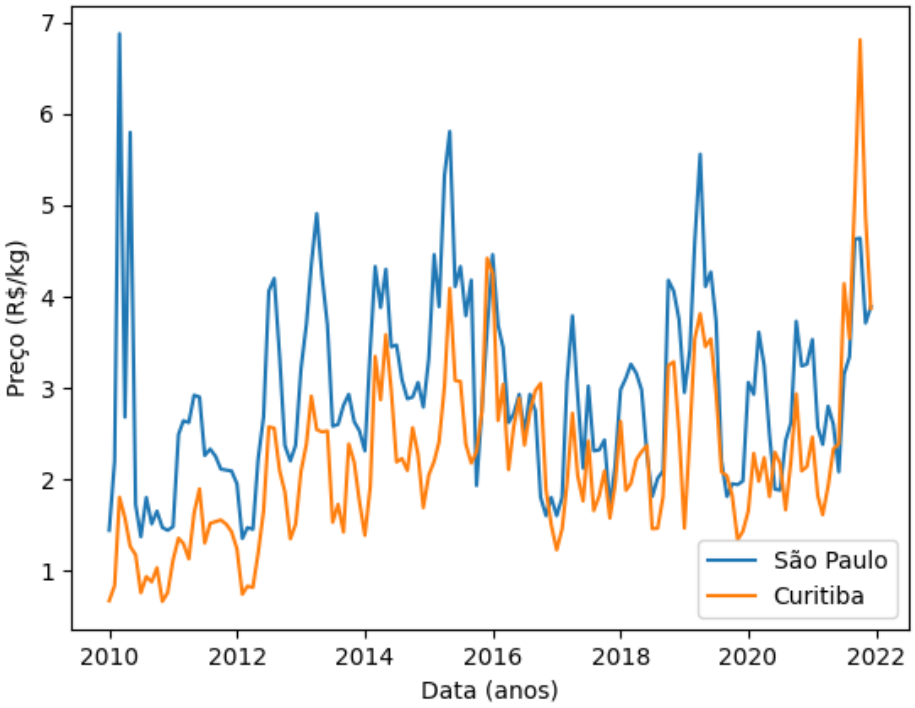
É possível observar que, de maneira geral, o IPCA apresentou variações ao longo dos anos, com momentos de elevação e queda. Em 2015, por exemplo, a média do IPCA foi de 0,951, valor bastante elevado em relação aos demais anos. De outro modo, em 2020, a média foi de 0.174, o que indica uma desaceleração significativa nos preços.

A quantidade de tomate comercializada em São Paulo aumenta entre 2010 e 2014, com um pico em janeiro de 2014. Em seguida, houve uma queda até meados de 2015, seguida por uma recuperação gradual. Após um pico em 2016, a quantidade comercializada diminuiu novamente até 2017 e permanece relativamente estável desde então. Em Curitiba, observa-se um padrão semelhante, porém, em vez de uma queda abrupta, a quantidade comercializada permaneceu relativamente estável até 2016, quando começou a diminuir gradualmente. A quantidade comercializada em São Paulo tem média maior do que em Curitiba, mas também uma maior variabilidade, como indicado pelo desvio padrão maior. Além disso, parece haver alguma flutuação na média ao longo do tempo, mas sem uma tendência clara.

Para visualizar melhor os dados de preços do tomate comercializado nas duas cidades, foi plotado o gráfico do Gráfico 3, que mostra a evolução ao longo dos anos.

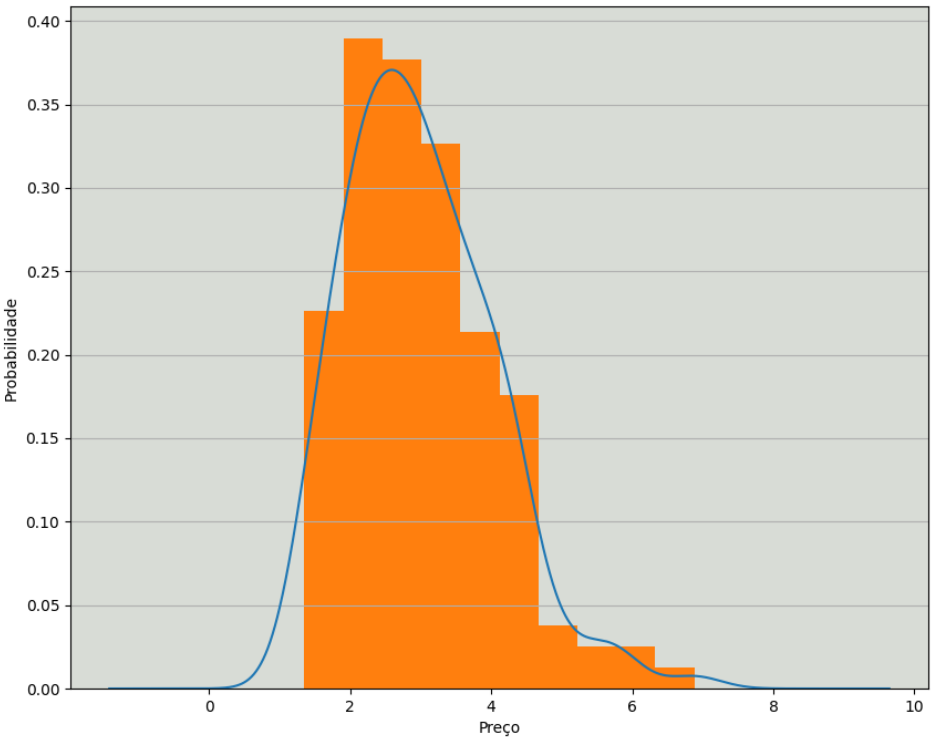
Os Gráficos 4 e 5 apresentam os histogramas dos preços de tomate nas duas cidades, com a probabilidade de ocorrência de cada intervalo.

Gráfico 3 - Evolução dos preços de tomate no atacado em São Paulo e Curitiba entre os anos de 2010 e 2021



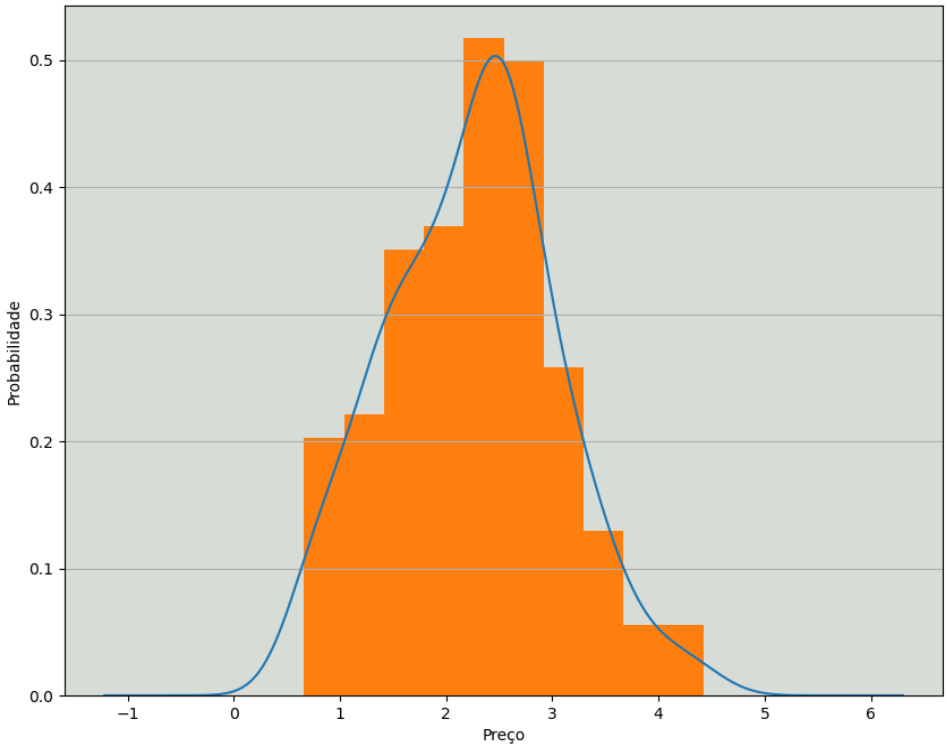
Fonte: O autor

Gráfico 4 - Histograma de probabilidade de ocorrência de preços no atacado de tomate no Ceagesp



Fonte: O autor

Gráfico 5 - Histograma de probabilidade de ocorrência de preços no atacado de tomate no Ceasa/PR



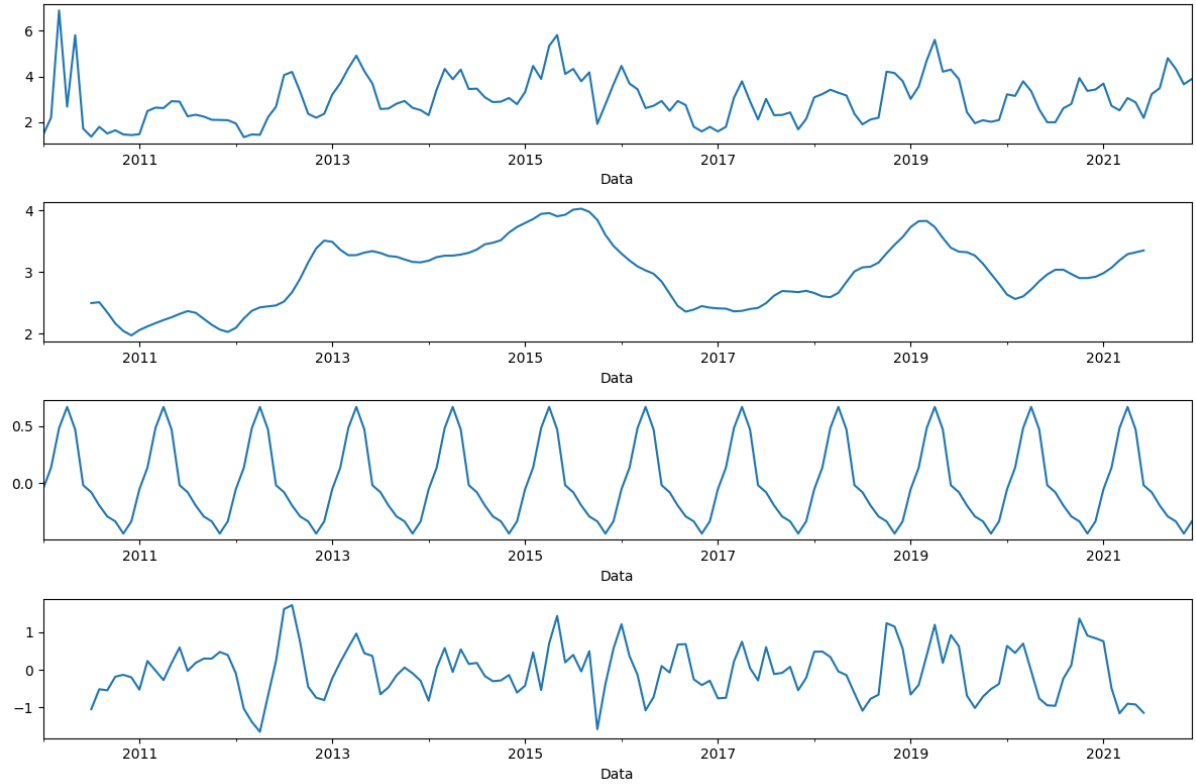
Fonte: O autor

Em São Paulo, os preços se concentram entre 1,5 e 5 Reais, enquanto em Curitiba ficam entre 1 e 3,5 Reais.

A curva de densidade de preços em São Paulo possui uma distribuição dos dados que não se assemelha a normal, dada a assimetria à direita e uma cauda longa. No gráfico de densidade em Curitiba, não há evidências claras de uma distribuição não normal. A curva parece simétrica e não há indícios de caudas longas ou assimetria pronunciada.

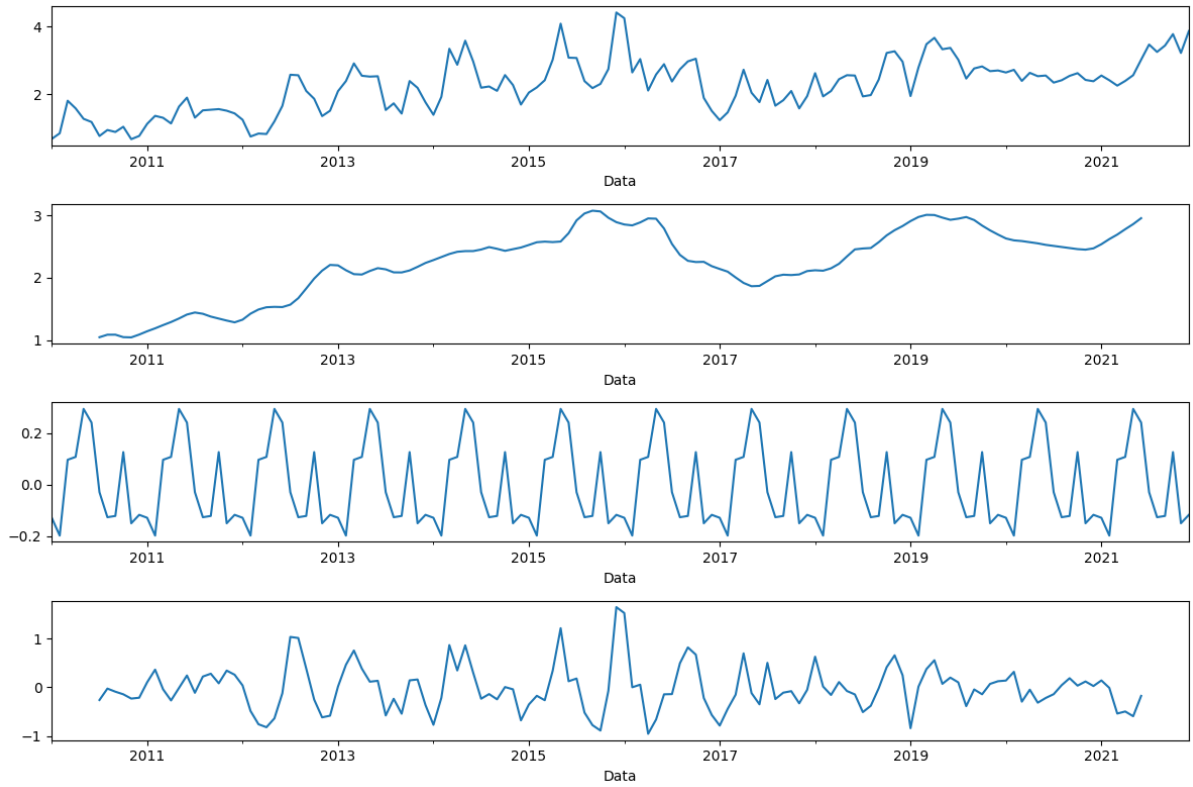
A decomposição das séries de preços em suas componentes de tendência, sazonalidade e resíduos é apresentada nos Gráficos 6 e 7.

Gráfico 6 - Decomposição para a série de preços de tomate em São Paulo



Fonte: O autor

Gráfico 7 - Decomposição para a série de preços de tomate em Curitiba



Fonte: O autor

Nota: Para os Gráficos 6 e 7 as decomposições estão na ordem descendente: dados originais, tendência, sazonalidade e resíduos.

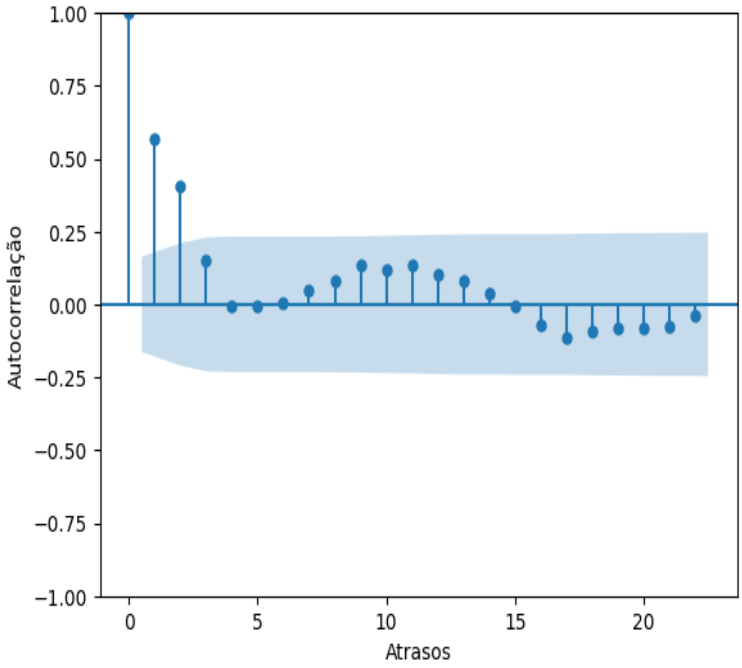
Nos gráficos, é possível observar que os preços do tomate apresentaram uma tendência de alta ao longo do período analisado, com algumas oscilações. Além disso, é possível notar que a sazonalidade é consistente, com valores maiores no primeiro semestre do ano, com pico observado entre os meses de maio e junho, seguido de uma queda gradual nos meses seguintes. Esse padrão pode ser explicado pelo fato de que a produção de tomate é afetada pelo clima, sendo que a colheita é mais abundante durante o verão (CONAB, 2019).

Ao longo dos anos, São Paulo apresentou preços mais elevados em relação a Curitiba, o que pode estar relacionado à maior demanda na capital paulista e aos custos de transporte.

Além disso, pode-se notar que os preços do tomate em Curitiba apresentaram uma maior variação ao longo do tempo, com um coeficiente anual médio de 45%, em comparação com o coeficiente de variação médio de 34% em São Paulo. Assim, os preços em Curitiba são mais voláteis, o que pode ser devido a fatores locais, como a oferta.

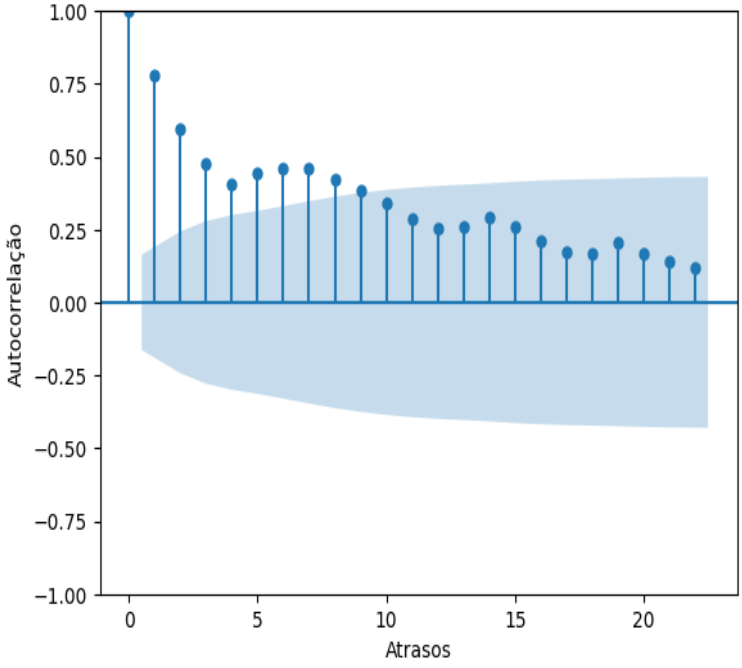
Os Gráficos 8 e 9 apresentam os gráficos de autocorrelação para valores de preço de tomate em São Paulo e Curitiba, respectivamente. O intervalo de confiança aplicado é de 95%.

Gráfico 8 - Autocorrelação para a série de preços do tomate no Ceagesp



Fonte: O autor

Gráfico 9 - Autocorrelação para a série de preços do tomate no Ceasa/PR



Fonte: O autor

No gráfico de autocorrelação, a série temporal para preços de tomate no Ceasa de São Paulo possui autocorrelação significativa para os três primeiros atrasos e diminui à medida

que o atraso aumenta, o que sugere uma tendência de decaimento exponencial. Isso é uma indicação de que a série temporal é provavelmente estacionária.

Ao analisar o gráfico para preços em Curitiba, nota-se que há uma correlação positiva até o atraso 9, com maior consistência até o terceiro. Após isso, há uma diminuição gradual da correlação, com flutuações em torno da média. Outro aspecto a ser observado é que as barras verticais do gráfico diminuem à medida que o atraso aumenta, indicando que a correlação é menos confiável em períodos mais distantes. Isso pode ser devido à influência de outros fatores que afetam a série no longo prazo, tornando a correlação menos significativa.

5.1 ANÁLISE DE ESTACIONARIEDADE

Para testar a estacionariedade das séries temporais utilizadas neste estudo, foi aplicado o teste Dickey-Fuller Aumentado (ADF). Os resultados para São Paulo e Curitiba são apresentados nas Tabelas 6 e 7.

Tabela 6 - Resultado da aplicação do teste de estacionariedade ADF para as variáveis em São Paulo

Variável	**Estatística ADF**	**p-value**	**Valor crítico 1%**	**Valor crítico 5%**	**Valor crítico 10%**
Quantidade comercializada	-1,46	0,55	-3,48	-2,88	-2,58
Precipitação	-2,70	0,07	-3,48	-2,88	-2,58
Temperatura	-2,14	0,23	-3,48	-2,88	-2,58
IPCA	-6,01	0,00	-3,48	-2,88	-2,58
Atividade do comércio	0,99	0,99	-3,48	-2,88	-2,58
Dólar	-0,34	0,92	-3,48	-2,88	-2,58
Preço Ureia	4,44	1,00	-3,48	-2,88	-2,58
Preços ao produtor	-6,06	0,00	-3,48	-2,88	-2,58
Diesel	1,49	1,00	-3,48	-2,88	-2,58
Preço do tomate	-4,83	0,00	-3,48	-2,88	-2,58

Fonte: O autor

Tabela 7 - Resultado da aplicação do teste de estacionariedade ADF para as variáveis em Curitiba

Variável	ADF Statistic	p-value	Valor crítico 1%	Valor crítico 5%	Valor crítico 10%
Quantidade comercializada	-3,78	0,00	-3,48	-2,88	-2,58
Precipitação	-10,95	0,00	-3,48	-2,88	-2,58
Temperatura	-2,13	0,23	-3,48	-2,88	-2,58
IPCA	-6,01	0,00	-3,48	-2,88	-2,58
Atividade do comércio	0,04	0,96	-3,48	-2,88	-2,58
Dólar	-0,34	0,92	-3,48	-2,88	-2,58
Preço Ureia	1,02	0,99	-3,48	-2,88	-2,58
Preços ao produtor	-2,06	0,26	-3,48	-2,88	-2,58
Diesel	1,80	1,00	-3,48	-2,88	-2,58
Preço do tomate	-1,93	0,32	-3,48	-2,88	-2,58

Fonte: O autor

Os resultados mostraram que a quantidade comercializada apresentou resultados diferentes entre as duas cidades. Enquanto em Curitiba a série se mostrou estacionária, em São Paulo não foi possível rejeitar a hipótese nula de não estacionariedade. Isso pode indicar que variações sazonais afetam de forma diferente as vendas em cada cidade.

Por outro lado, a precipitação apresentou resultados negativos para ambas as cidades, indicando que a série é estacionária. Esse resultado é esperado, pois a precipitação é uma variável climática que tende a apresentar média e variância constantes ao longo do tempo (MILLY et al., 2008). Porém, a variável temperatura, apresentou comportamento não-estacionário.

O IPCA, que mede a inflação, apresentou estatísticas ADF negativas e p-valores baixos, o que sugere uma série estacionária. As flutuações na inflação são devidas principalmente a choques temporários na economia e não a uma tendência de longo prazo (BLANCHARD e SHEEN, 2012).

Em relação à atividade do comércio e preço do diesel, os valores-p próximos de um indicam que as séries são não-estacionárias. Esse resultado pode ser explicado pelo fato de que a atividade do comércio pode variar consideravelmente ao longo do tempo, dependendo de fatores econômicos e sazonais (MORETTIN; TOLOI, 2018).

Em relação ao preço da ureia, os resultados mostraram uma estatística ADF de 4,44 em São Paulo e 1,02 em Curitiba, indicando que a série não é estacionária em nenhuma das cidades.

O preço do tomate apresentou resultados discrepantes entre as duas cidades. A série se mostrou não estacionária em Curitiba, mas estacionária em São Paulo. Essa diferença pode ser explicada por variações sazonais que afetam de forma diferente a oferta e a demanda desse produto em cada cidade (GALLO, 2007).

5.2 CORRELAÇÃO ENTRE AS VARIÁVEIS

As correlações apresentadas nas Tabelas 8 e 9 mostram os resultados da análise de correlação linear entre as variáveis objeto deste estudo.

Tabela 8 - Matriz de correlação entre as variáveis cotadas para integrar os modelos de predição em São Paulo

	Qtd	Pct	Temp	IPCA	Atv	Dolar	Ureia	Prc_pr	Diesel	Prec_to
Qtd	1,00									
Pct	0,27	1,00								
Temp	0,38	0,65	1,00							
IPCA	0,33	0,31	0,37	1,00						
Atv	-0,36	0,00	0,08	0,15	1,00					
Dolar	-0,37	-0,04	0,02	0,09	0,87	1,00				
Ureia	-0,30	-0,09	-0,07	0,21	0,79	0,76	1,00			
Prc_pr	-0,07	-0,13	-0,04	0,02	0,22	0,29	0,28	1,00		
Diesel	-0,44	-0,05	0,02	0,10	0,88	0,90	0,86	0,27	1,00	
Prec_to	-0,12	-0,06	0,02	0,18	0,15	0,17	0,20	0,40	0,21	1,00

Fonte: O autor

Tabela 9 - Matriz de correlação entre as variáveis cotadas para integrar os modelos de predição em Curitiba

	Qtd	Pct	Temp	IPCA	Atv	Dolar	Ureia	Prc_pr	Diesel	Prec_to
Qtd	1,00									
Pct	-0,22	1,00								
Temp	-0,16	0,33	1,00							
IPCA	-0,18	0,25	0,34	1,00						
Atv	0,14	0,01	0,11	0,13	1,00					
Dolar	0,17	-0,02	0,03	0,09	0,86	1,00				
Ureia	0,21	-0,01	-0,06	0,33	0,57	0,56	1,00			
Prc_pr	-0,07	0,06	0,00	0,15	0,50	0,53	0,52	1,00		
Diesel	0,22	-0,02	0,04	0,12	0,89	0,89	0,71	0,61	1,00	
Prec_to	-0,19	0,04	-0,10	0,16	0,62	0,59	0,43	0,82	0,64	1,00

Fonte: O autor

Notas: Para as Tabelas 8 e 9 considerar as abreviaturas: quantidades mensais de tomate comercializado (Qtd), volumes mensais de precipitação (pct), médias mensais de temperatura(Temp), taxas mensais do Índice de Preços ao Consumidor (IPCA), valores mensais para atividade do comércio (Atv), médias mensais de preço do Dólar (Dolar), médias mensais de preço do fertilizante ureia (Ureia), médias mensais de preços pegos aos produtores de tomate (Pcr_pr), médias mensais de preços praticados para o óleo diesel (Diesel), médias mensais de preços do tomate nos Ceasas (Prec_to).

Na cidade de São Paulo, podemos observar que a correlação entre a quantidade comercializada e o preço do tomate é negativa, ou seja, quando a quantidade comercializada aumenta, o preço do tomate tende a diminuir. De modo diferente, em Curitiba, a correlação entre a quantidade comercializada e o preço do tomate é próxima de zero, indicando que não há uma relação muito forte entre essas variáveis.

As variáveis que apresentam uma correlação mais forte com o preço do tomate para São Paulo são o preço do diesel, o preço pago ao produtor e o valor da ureia. No Paraná, além das duas primeiras, o índice de atividade do comércio tem valor significativo nessa análise.

Em São Paulo, a quantidade de tomate comercializada apresenta uma correlação positiva moderada com a precipitação total, a temperatura e o IPCA. Porém, há uma correlação negativa moderada entre o total vendido e a atividade do comércio, o dólar, a ureia e o preço do diesel. Isso pode ser explicado pelo fato de que, com o dólar mais forte, preços mais elevados do diesel e da ureia, e a atividade comercial menos intensa podem afetar negativamente o comércio de tomate.

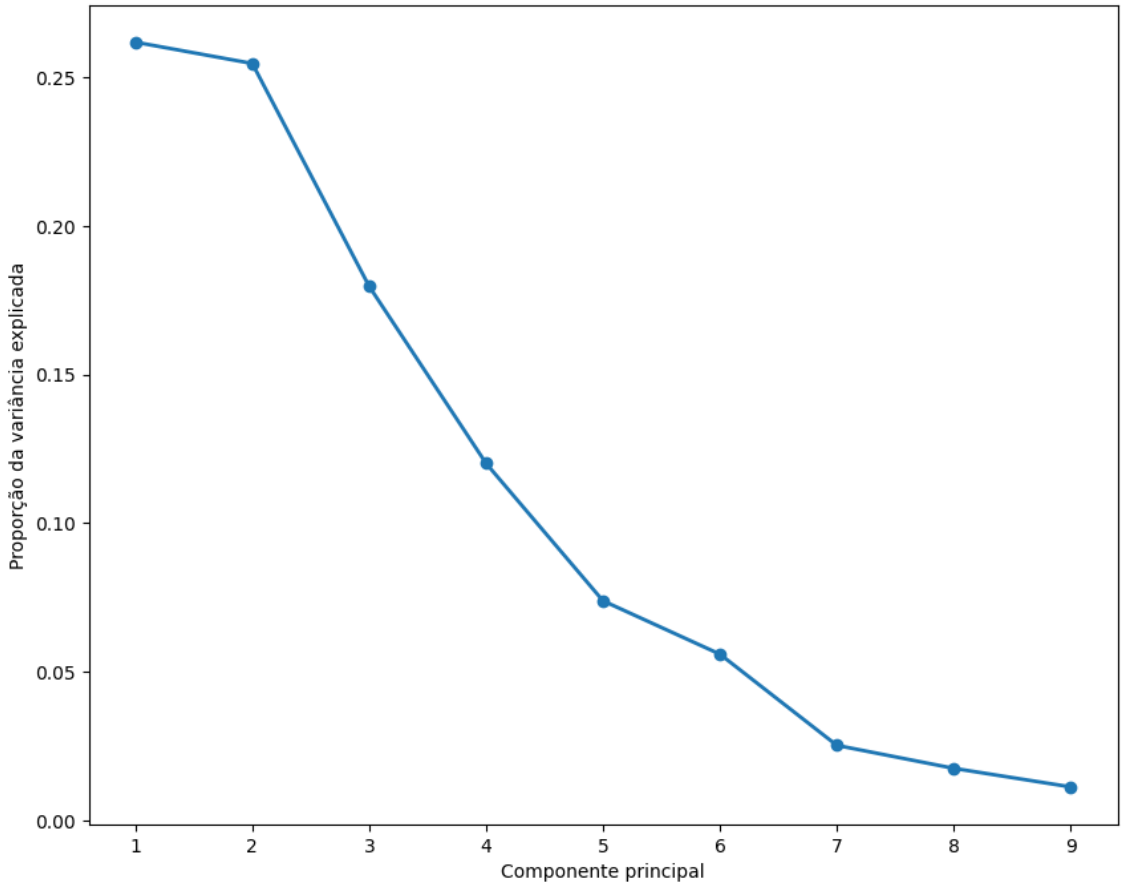
A correlação negativa fraca entre a quantidade comercializada e os preços pagos ao produtor sugere que, à medida que os preços aumentam, a quantidade comercializada diminui, mas de forma discreta. Isso é válido para os dois locais.

Em Curitiba a quantidade comercializada apresenta uma correlação negativa fraca com a precipitação total, a temperatura e o IPCA. Por outro lado, é possível observar correlação positiva moderada entre a quantidade e a atividade comercial.

5.3 REDUÇÃO DA DIMENSIONALIDADE

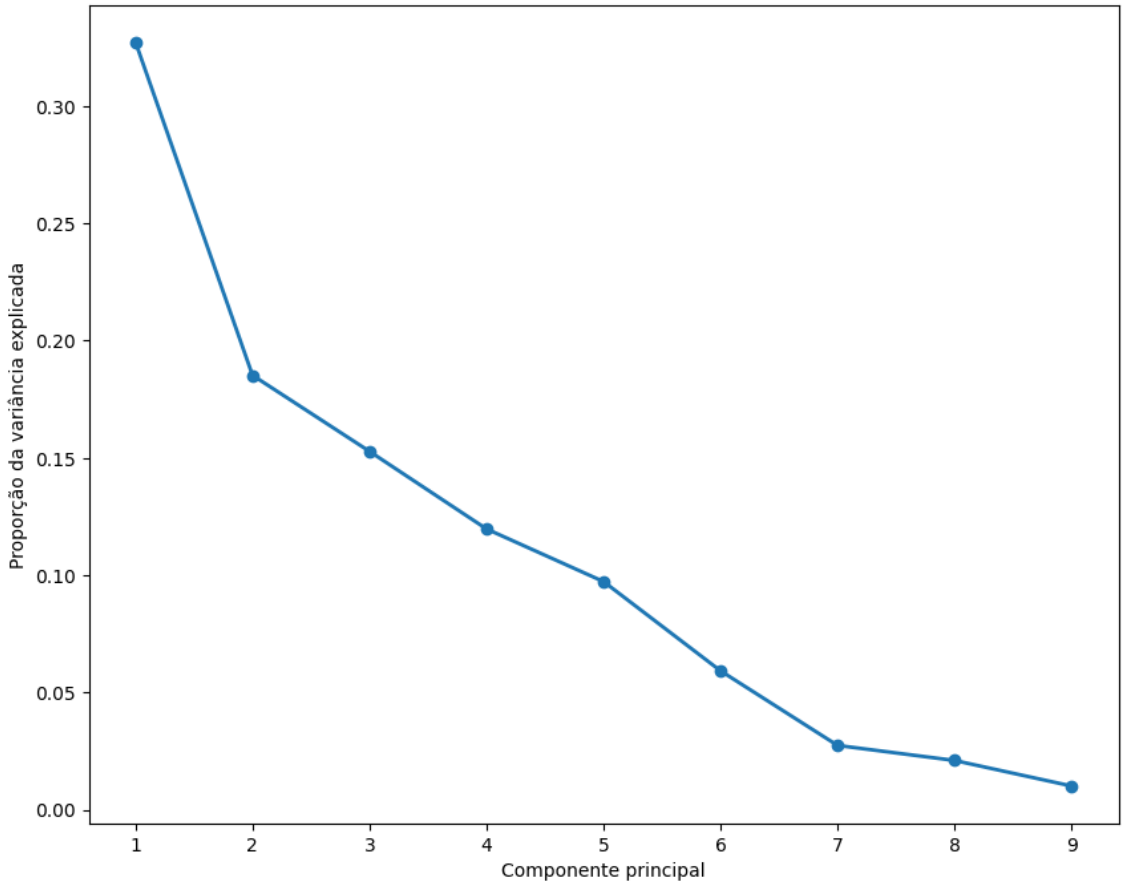
Como pode ser observado nos gráficos das Ilustrações 10 e 11, o número de componentes principais necessários para explicar pelo menos 80% da variância, tanto no conjunto de dados para São Paulo, como para Curitiba, é 5. Essas séries têm cargas fatoriais relativamente altas, o que sugere que são importantes para a explicação da variabilidade total dos dados.

Gráfico 10 - Proporção da variância explicada em cada componente principal para São Paulo



Fonte: O autor

Gráfico 11 - Proporção da variância explicada em cada componente principal para Curitiba



Fonte: O autor

As variáveis selecionadas nas duas bases foram as mesmas, e são elas: atividade do comércio, precipitação, temperatura, IPCA e quantidade de tomate comercializada.

As variáveis selecionadas pela técnica de ACP foram utilizadas para a análise comparativa dos métodos candidatos deste estudo.

5.4 ANÁLISE DOS MODELOS

Esta seção apresenta os resultados de cada modelo objeto deste estudo para as simulações nos conjuntos de dados.

5.4.1 ARIMA

Os valores para as métricas de desempenho, RMSE e MAPE, assim como as equações de saída do método ARIMA para São Paulo e Curitiba são apresentadas na Tabela 10.

Tabela 10 - Resultados para as métricas de desempenho RMSE e MAPE e o modelo ARIMA

Local	RMSE	MAPE	Modelo de Saída
São Paulo	0,22	0,43	y(t) = 0,29 + 0,47y(t-1) + 0,43y(t-2) - 0,35*e(t-1)
Curitiba	0,14	0,24	y(t) = 0,49 + 0,68*y(t-1) - 0,20*y(t-2) - 0,90*e(t-1)

Fonte: O autor

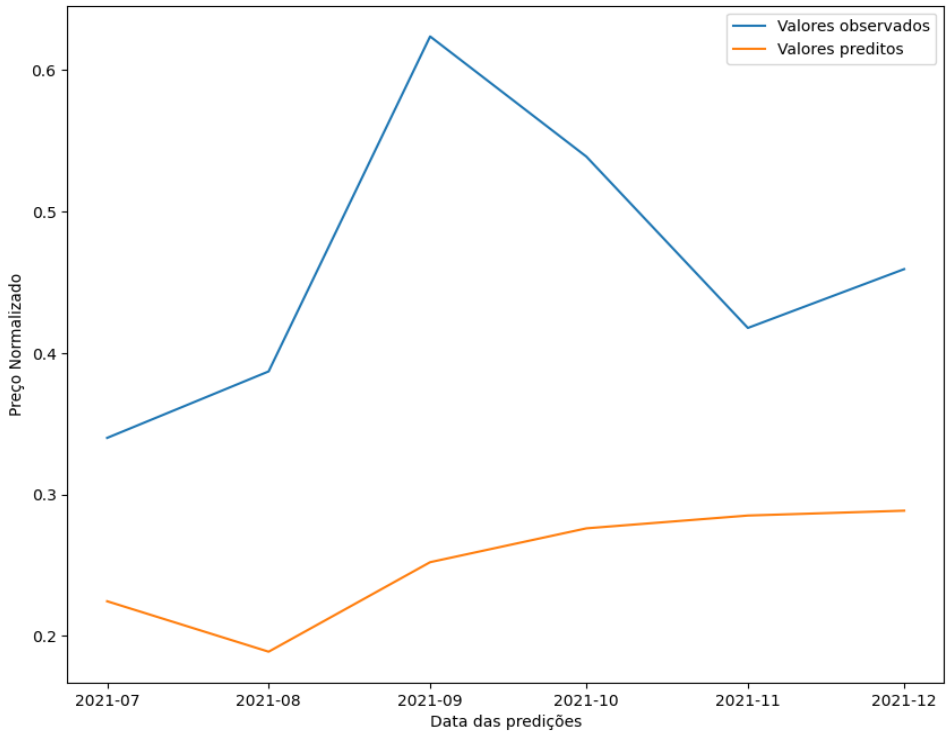
Como o método de médias autorregressivas é univariado, apenas a série de preços do tomate foi utilizada na modelagem.

Os erros para as previsões realizadas para os dados de Curitiba foram menores do que aqueles obtidos para São Paulo. Para as duas localidades, os modelos apresentaram baixa capacidade de previsão, com valores para métrica MAPE acima de 20%.

Nas equações, y(t) representa o preço do tomate no tempo t. Como definido na modelagem, o preço do tomate no tempo t é influenciado pelos preços do tomate nos dois períodos de tempo anteriores, (t-1 e t-2). A equação para ambos os locais mostram um peso maior para o preço no período anterior, (t-1).

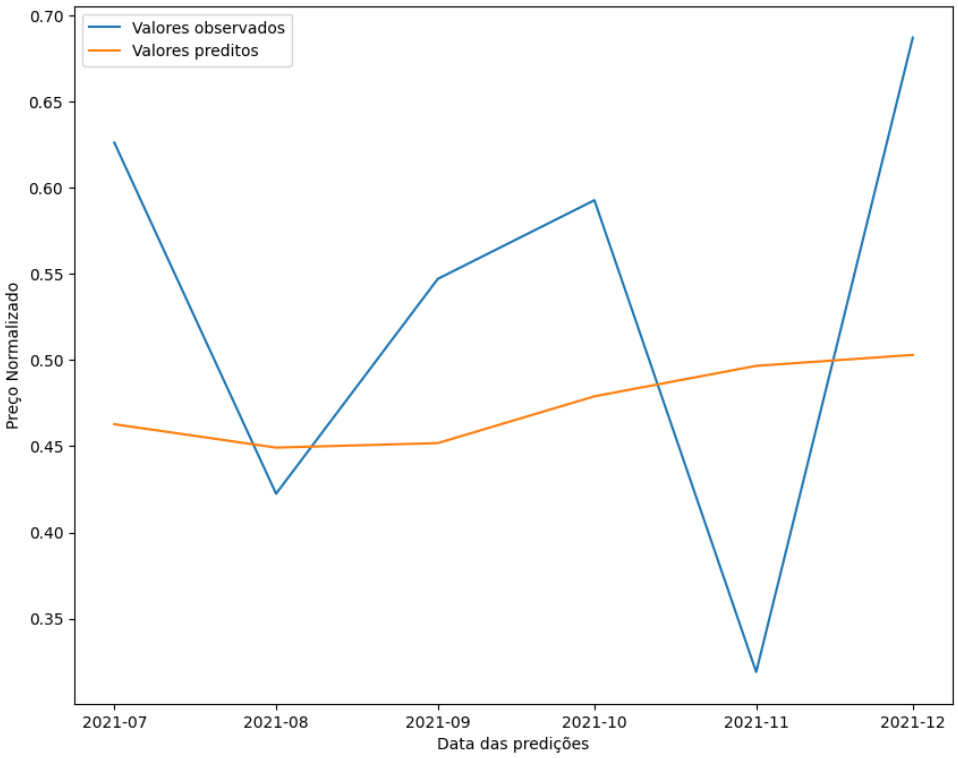
Os Gráficos 12 e 13 apresentam os valores de predição versus aqueles observados na amostra.

Gráfico 12 - Séries de valores reais e de teste para previsão no Ceagesp pelo método ARIMA



Fonte: O autor

Gráfico 13 - Séries de valores reais e de teste para previsão no Ceasa/PR pelo método ARIMA



Fonte: O autor

Pode-se observar que o modelo não foi capaz de capturar as tendências e variações da série, com baixa correspondência entre os valores previstos e observados.

5.4.2 SARIMA

Assim como o método ARIMA, o SARIMA é univariado, apenas as séries de preços do tomate foram utilizadas na modelagem.

Os valores para as métricas de desempenho, RMSE e MAPE, assim como as equações de saída do método SARIMA para São Paulo e Curitiba são apresentadas na Tabela 11.

Tabela 11 - Resultados para as métricas de desempenho RMSE e MAPE e o modelo SARIMA

Local	RMSE	MAPE	Modelo de Saída
São Paulo	0,26	0,51	y(t) = 0.44 + 0.41*y(t-1) - 0.27*y(t-2) - 0.82*e(t-1) + 0.02*x(t)
Curitiba	0,16	0,30	y(t) = 0,03 – 0,25*y(t-1) - 0.86*y(t-2) + 0.06*e(t-1) + 0.06*x(t)

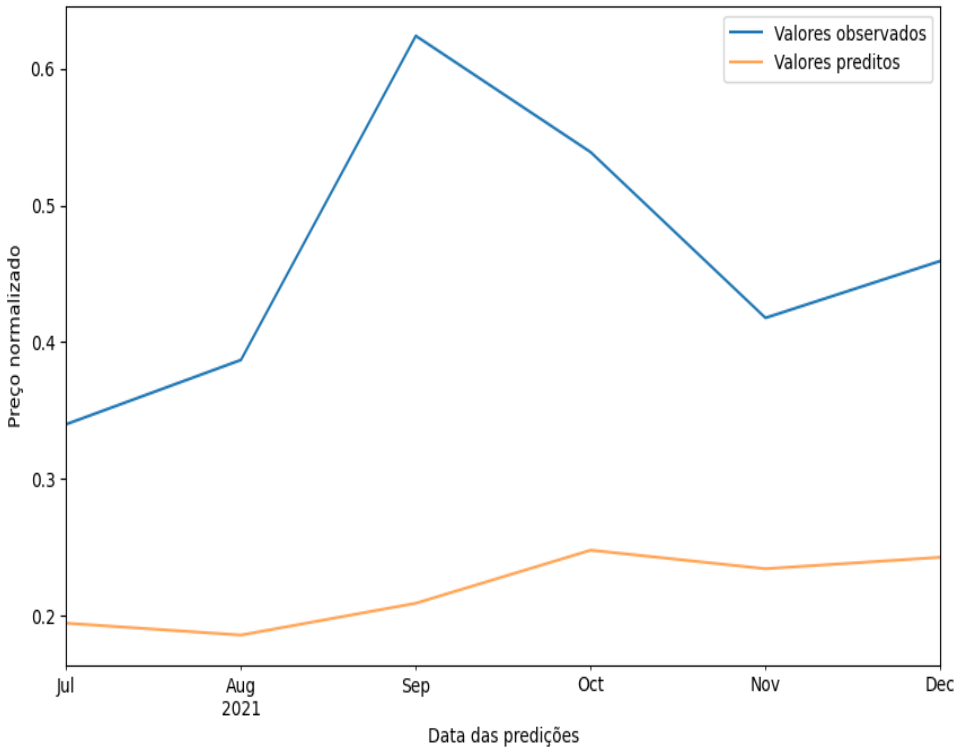
Fonte: O autor

A partir dos resultados apresentados, podemos concluir que o modelo SARIMA obteve desempenho melhor na previsão de preços de tomate no atacado para a unidade de Curitiba, mas a precisão para os dois locais é classificada como baixa.

O modelo para Curitiba apresenta coeficientes negativos para os valores anteriores de tomate, o que indica tendência de queda dos preços. Esse modelo também apresentou coeficiente mais significativo para o termo y(t-2). Com isso, as informações históricas dos preços de tomate em Curitiba são mais importantes para prever os preços futuros do que as variáveis exógenas, que são contempladas pelo termo de erro.

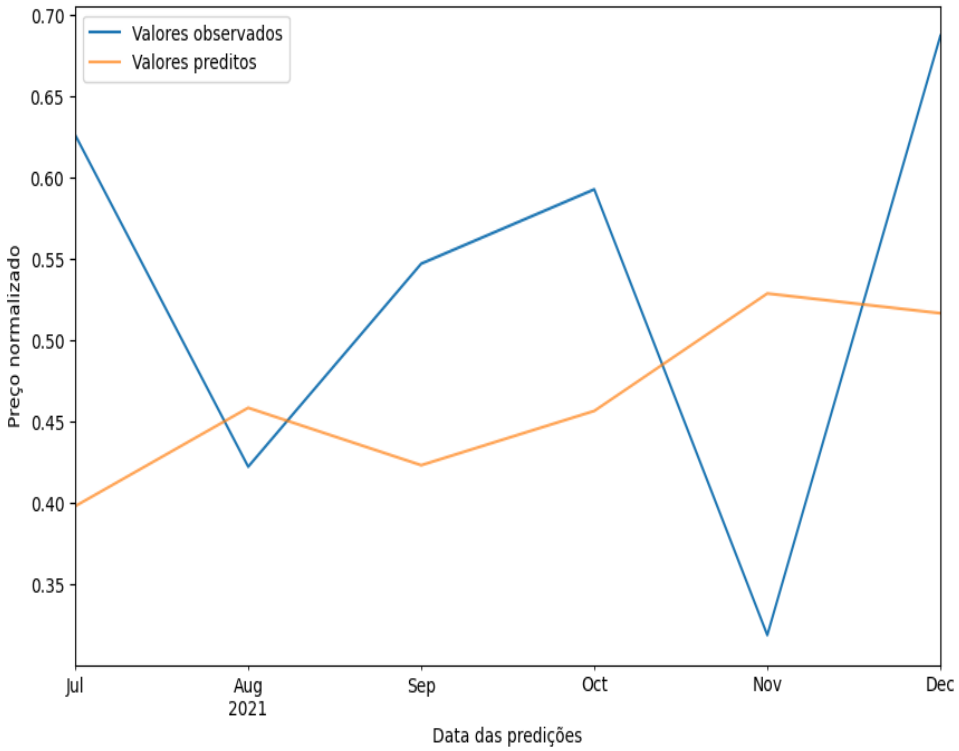
Os Gráficos 14 e 15 mostram a comparação entre os valores observados e as previsões feitas pelo modelo para os locais analisados.

Gráfico 14 - Séries de valores reais e de teste para previsão no Ceagesp pelo método SARIMA



Fonte: O autor

Gráfico 15 - Séries de valores reais e de teste para previsão no Ceasa/PR pelo método SARIMA



Fonte: O autor

Observa-se que o modelo conseguiu capturar de forma mais precisa as variações sazonais dos preços de tomate em Curitiba. Para São Paulo, o gráfico indica que as flutuações, movimentos de alta e baixa, não foram captadas.

5.4.3 ARIMAX

Ao contrário dos métodos ARIMA e SARIMA, que são univariados, o ARIMAX é multivariado. Para essa modelagem foram utilizadas as variáveis selecionadas pelo método de análise de componentes principais.

Os valores para as métricas de desempenho, RMSE e MAPE, assim como as equações de saída do modelo ARIMAX para São Paulo e Curitiba são apresentadas na Tabela 12.

Tabela 12 - Resultados para as métricas de desempenho RMSE e MAPE e modelo ARIMAX

Local	**RMSE**	**MAPE**	**Modelo de Saída**
São Paulo	0,21	0,41	y(t) = 0.29 + 0.45 * y(t-1) + 0.45 * y(t-2) + 0.01 * exog1(t) - 0.01 * exog2(t) - 0.03 * exog3(t) - 0.02 * exog4(t) - 0.00 * exog5(t) - 0.36 * e(t-1) + 0.02 * e(t-2)
Curitiba	0,09	0,17	y(t) = 0.49 + 0.66 * y(t-1) - 0.21 * y(t-2) + 0.01 * exog1(t) + 0.02 * exog2(t) - 0.04 * exog3(t) + 0.01 * exog4(t) + 0.02 * exog5(t) - 0.88 * e(t-1) + 0.02 * e(t-2)

Fonte: O autor

Notas: Considerar exog1 (IPCA), exog2 precipitação, exog3 (temperatura), exog4 (quantidade comercializada), exog5 (índice de atividade do comércio).

O método ora analisado, assim como os anteriores, apresenta melhor desempenho para a previsão de preços do tomate no atacado para a série de dados do Ceasa/PR, para a qual a capacidade de previsão é razoável.

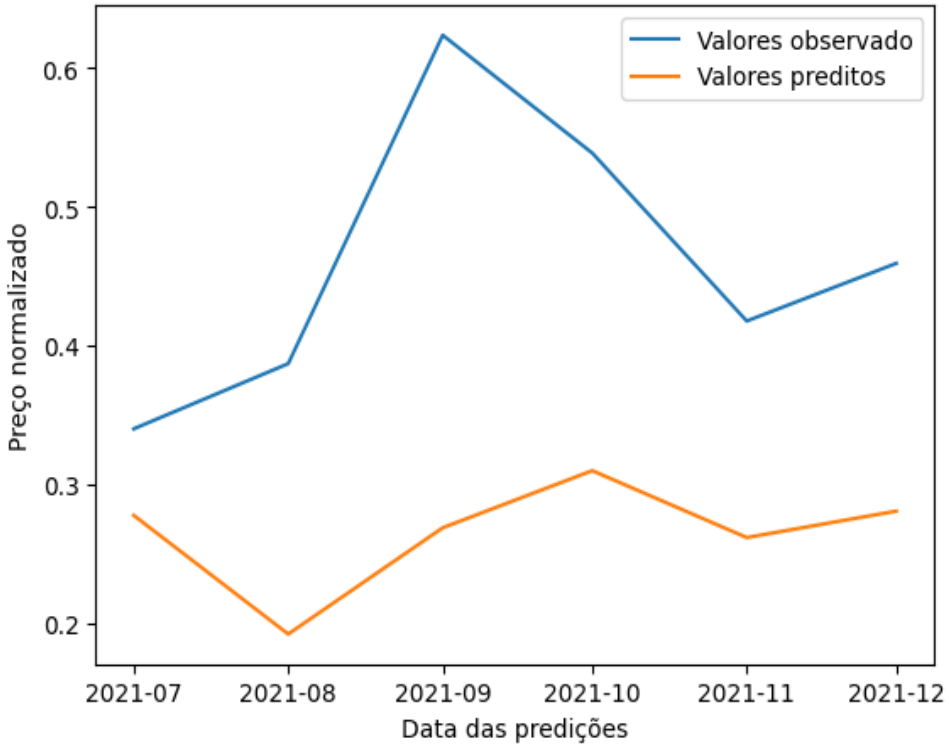
No modelo para São Paulo, os coeficientes dos termos autorregressivos são positivos, o que significa que valores anteriores têm um impacto positivo no preço atual. Para a capital do Paraná, o coeficiente para y(t-1) comparativamente maior e positivo denota uma influência mais significativa dos preços imediatamente anteriores. No entanto, o coeficiente para y(t-2) é negativo, o que sugere uma influência negativa dos valores mais antigos.

O índice de atividade do comércio não foi relevante para o modelo de São Paulo. Em Curitiba, com exceção da temperatura, as demais variáveis apresentam correlação positiva com a série de preços do tomate.

Vale destacar que, para os dois locais, a variável exógena mais relevante para os modelos foi a temperatura, com maior peso em Curitiba.

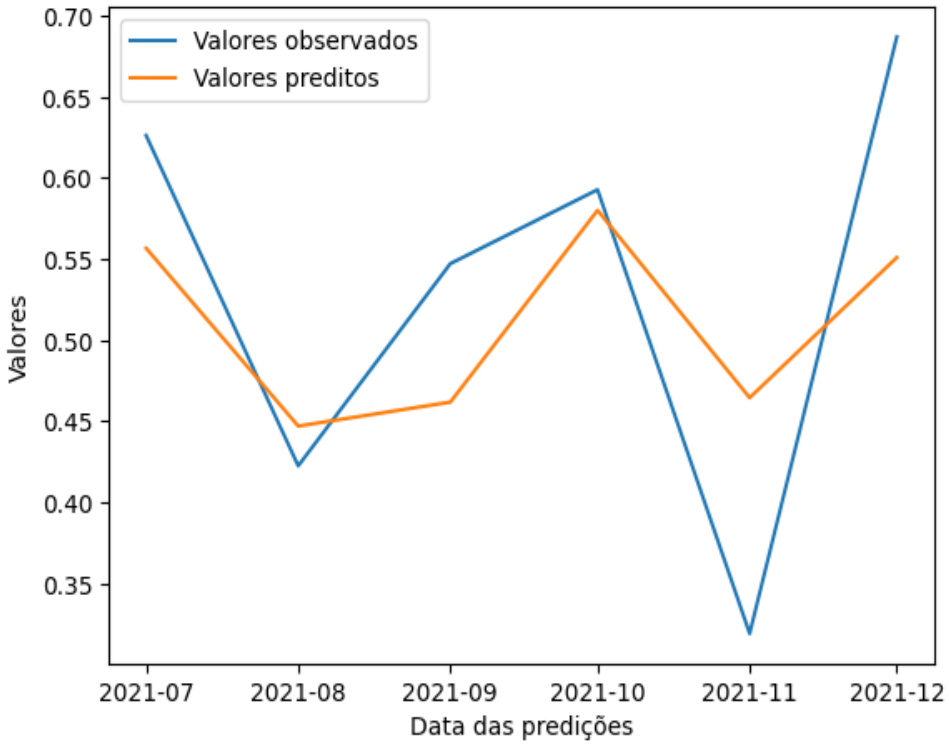
Nos gráficos das Ilustrações 16 e 17 são apresentadas as comparações entre os valores observados e as previsões feitas pelos modelos.

Gráfico 16 - Séries de valores reais e de teste para a previsão no Ceagesp pelo método ARIMAX



Fonte: O autor

Gráfico 17 - Séries de valores reais e de teste para a previsão no Ceasa/PR pelo método ARIMAX



Fonte: O autor

Os gráficos evidenciam o melhor desempenho do modelo sobre os dados de preço em Curitiba.

Dos três métodos estatísticos clássicos analisados, o ARIMAX é aquele com os menores erros de previsão. Isso sugere que, ao incluir variáveis exógenas, as previsões foram mais precisas.

5.4.4 SVR

Diferente dos métodos clássicos, a regressão vetorial de suporte não pode ser facilmente expressa em forma de equação, pois envolve uma etapa de treinamento em que os parâmetros são ajustados para minimizar a diferença entre as previsões e os valores observados. Esses parâmetros são determinados de maneira iterativa usando algoritmos de otimização complexos, o que torna difícil expressar o modelo em uma equação fechada (SMOLA; SCHÖLKOPF, 2004).

Embora não seja possível apresentar o modelo SVR em forma de equação, é possível interpretar como os fatores são usados pelo modelo e entender como eles afetam as previsões. Para isso, a Tabela 13 apresenta o valor de importância das variáveis em seus respectivos atrasos.

Tabela 13 - Importância das variáveis no modelo SVR para São Paulo e Curitiba

(continua)

Variável	Atraso	São Paulo	Curitiba
Índice de Atividade do Comércio	0	0,04	0,07
Índice de Atividade do Comércio	1	0,02	0,03
Índice de Atividade do Comércio	2	0,03	-0,04
Índice de Atividade do Comércio	3	0,03	0,02
IPCA	0	0,01	-0,03
IPCA	1	0,02	0,06
IPCA	2	-0,03	0,06
IPCA	3	-0,02	0,02
Precipitação	0	0,01	0,09
Precipitação	1	0,01	0,01
Precipitação	2	-0,02	0,07
Precipitação	3	-0,02	0,00
Preço do tomate	1	1,21	0,00

Tabela 13 - Importância das variáveis no modelo SVR para São Paulo e Curitiba

(conclusão)

Variável	Atraso	São Paulo	Curitiba
Preço do tomate	2	0,07	-0,01
Preço do tomate	3	0,25	-0,03
Quantidade comercializada	0	0,15	0,12
Quantidade comercializada	1	0,13	-0,04
Quantidade comercializada	2	0,01	-0,01
Quantidade comercializada	3	-0,04	0,00
Temperatura	0	0,06	0,07
Temperatura	1	0,07	-0,02
Temperatura	2	0,47	0,11
Temperatura	3	0,12	0,03

Fonte: O autor

As variáveis mais importantes para o modelo em São Paulo são o preço do tomate no mês anterior e a temperatura com dois períodos de atraso, com valores de 1,21 e 0,47, respectivamente. A primeira variável possui uma importância proporcionalmente alta, o que pode indicar superajuste aos dados de treinamento (LUNDBERG; LEE, 2017).

De forma oposta, os resultados para Curitiba indicam que os preços anteriores da série não possuem peso significativo para o modelo, sendo a quantidade comercializada no mês corrente a variável com maior peso e, assim, assim como para a capital do sudeste, esse modelo indica que a temperatura com atraso de dois meses também é relevante.

A Tabela 14 mostra os valores das métricas de desempenho RMSE e MAPE para o modelo SVR no Ceagesp e Ceasa/PR.

Tabela 14 - Resultado das métricas de desempenho RMSE e MAPE do modelo SVR

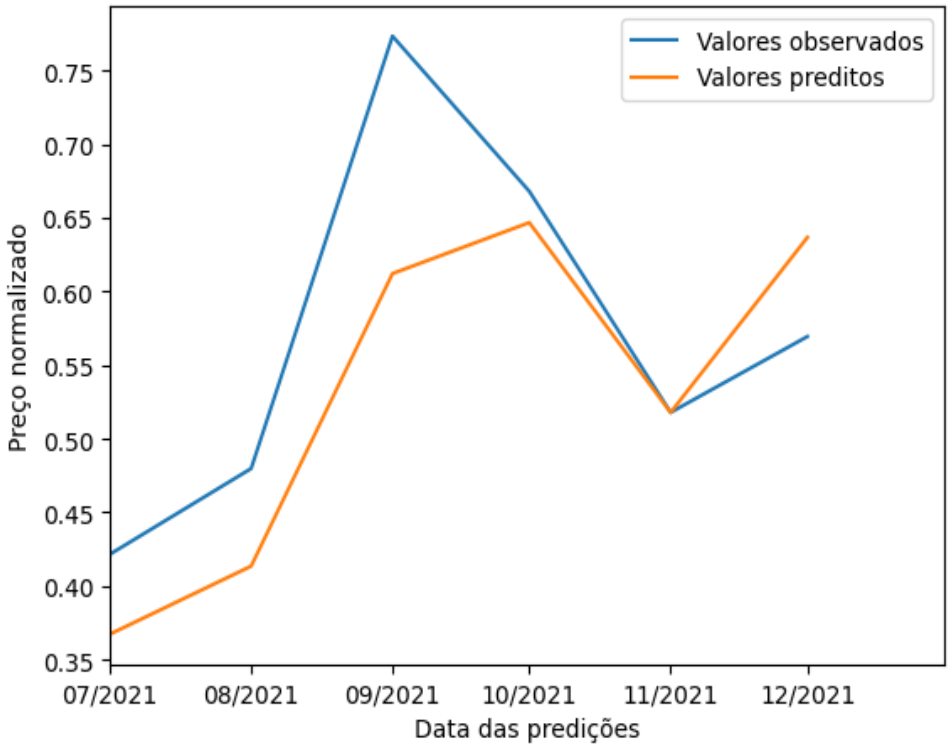
Local	RMSE	MAPE
São Paulo	0,08	0,10
Curitiba	0,08	0,12

Fonte: O autor

Os resultados indicam que o modelo SVR teve um desempenho semelhante nas duas cidades, com valores que indicam previsões próximas das reais e, com isso, o modelo pode ser classificado como de boa precisão para São Paulo e razoável para Curitiba.

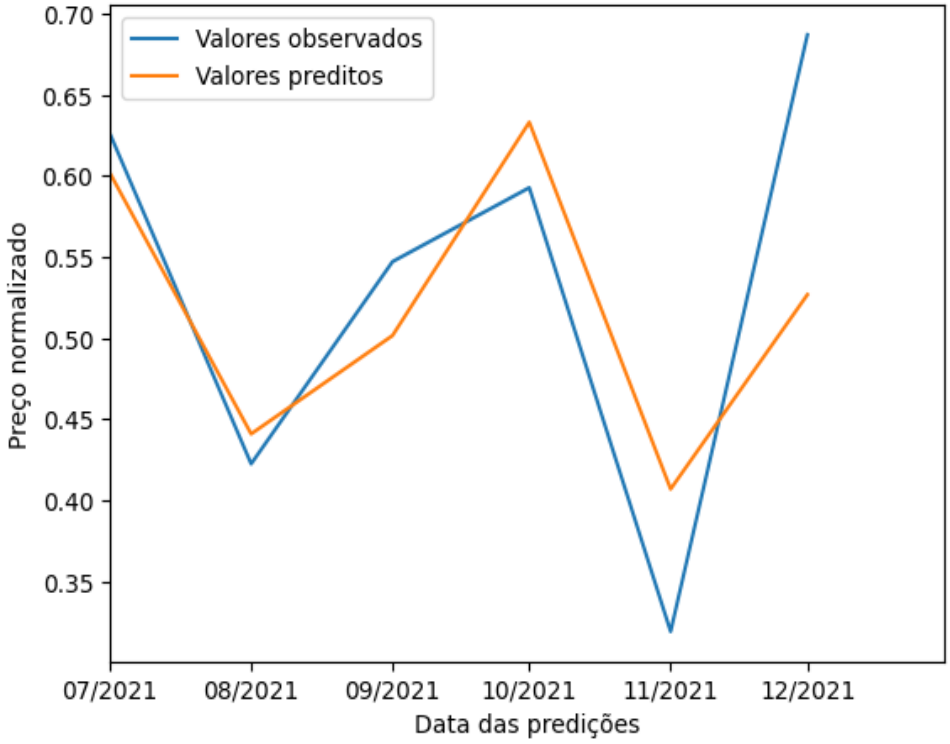
Os Gráficos 18 e 19 apresentam os gráficos com valores preditos pelo modelo SVR e os valores reais.

Gráfico 18 - Séries de valores reais e de teste para a previsão de preços no Ceagesp pelo método SVR



Fonte: O autor

Gráfico 19 - Séries de valores reais e de teste para a previsão de preços no Ceasa/PR pelo método SVR



Fonte: O autor

Para os dois locais alvos deste estudo, o modelo foi capaz de capturar oscilações do preço. Para São Paulo é possível observar que o movimento de alta da série foi subestimado, enquanto que para Curitiba as estimativas para fundos e vales foi mais precisa.

5.4.5 LSTM

O método LSTM não tem uma forma analítica que permita obter uma equação de saída, pois depende do aprendizado dos pesos da rede neural e do estado oculto dinâmico da rede (HOCHREITER; SCHMIDHUBER, 1997).

Para auxiliar na interpretação dos parâmetros e na análise dos pesos do modelo são apresentados na Tabela 15 os valores de importância para as variáveis e seus respectivos atrasos.

Tabela 15 - Importância das variáveis no modelo LSTM para São Paulo e Curitiba

(continua)

Variável	Atraso	São Paulo	Curitiba
Índice de Atividade do Comércio	0	31,40	16,93
Índice de Atividade do Comércio	1	5,65	-6,43
Índice de Atividade do Comércio	2	-13,30	1,98
Índice de Atividade do Comércio	3	8,93	6,21
IPCA	0	11,98	-25,34
IPCA	1	-3,81	-1,45
IPCA	2	6,73	9,58
IPCA	3	15,64	-18,70
Precipitação	0	-15,74	-14,32
Precipitação	1	3,32	-1,69
Precipitação	2	10,50	7,21
Precipitação	3	5,64	5,73
Preço do tomate	1	-24,60	-0,24
Preço do tomate	2	76,77	-1,15
Preço do tomate	3	-12,21	-9,13
Quantidade comercializada	0	6,67	8,11
Quantidade comercializada	1	-1,88	8,48
Quantidade comercializada	2	1,39	-1,56

Tabela 15 - Importância das variáveis no modelo LSTM para São Paulo e Curitiba

(conclusão)

Variável	Atraso	São Paulo	Curitiba
Quantidade comercializada	3	4,08	0,21
Temperatura	0	-11,69	74,73
Temperatura	1	-4,42	58,25
Temperatura	2	-8,48	-11,47
Temperatura	3	7,43	-5,95

Fonte: O autor

Na tabela apresentada, observa-se que as variáveis mais importantes para o modelo em São Paulo são preço do tomate com atraso de dois meses, índice de atividade do comércio no mês corrente e IPCA com atraso de três meses, com valores positivos relativamente altos.

Para Curitiba, a variável mais importante para o modelo é a temperatura nos meses corrente e com atraso de um período. Esse resultado é coerente ao se perceber que esse local apresenta estações do ano bem definidas, com verões moderadamente quentes e invernos frios (SILVA at al., 2019). É comum que a cidade experimente geadas e eventualmente temperaturas próximas ou abaixo de zero durante o inverno.

A Tabela 16 mostra os resultados das métricas RMSE e MAPE do modelo LSTM para previsão de preços de tomate no atacado em São Paulo e Curitiba.

Tabela 16 - Resultado das métricas de desempenho RMSE e MAPE do modelo LSTM

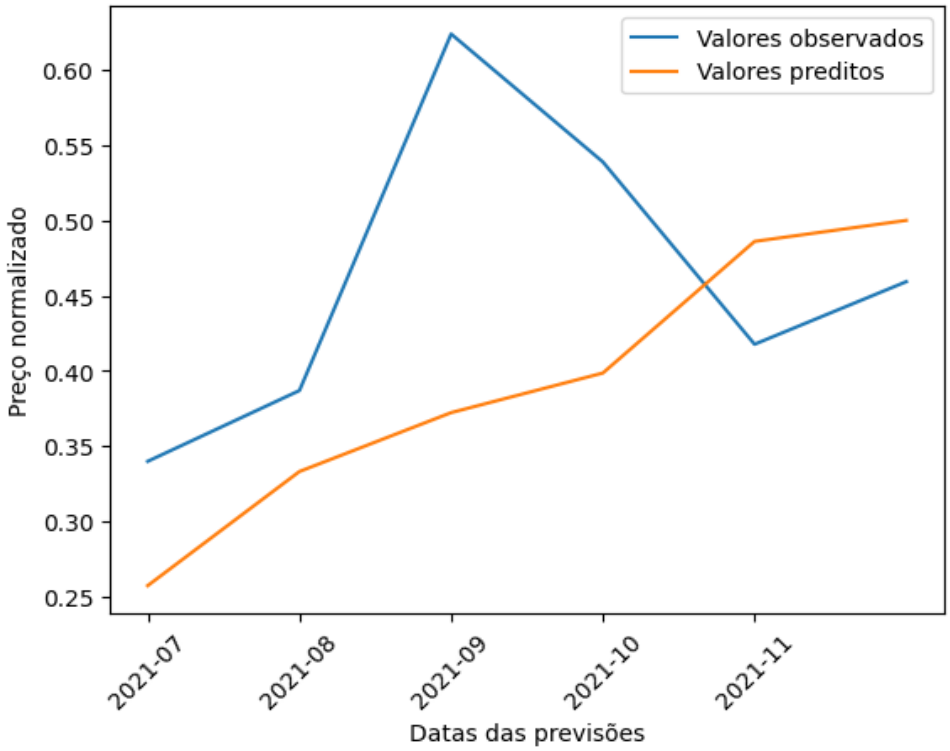
Local	RMSE	MAPE
São Paulo	0,12	0,23
Curitiba	0,12	0,22

Fonte: O autor

Os resultados mostram que não houve diferença entre os dois locais para os valores de RMSE e a diferença observada para MAPE é de 1%. Essa proximidade entre as métricas dos dois locais também foi observada no modelo SVR. Ainda, ao serem considerados resultados para a segunda métrica, o modelo é classificado com baixa precisão para as duas capitais.

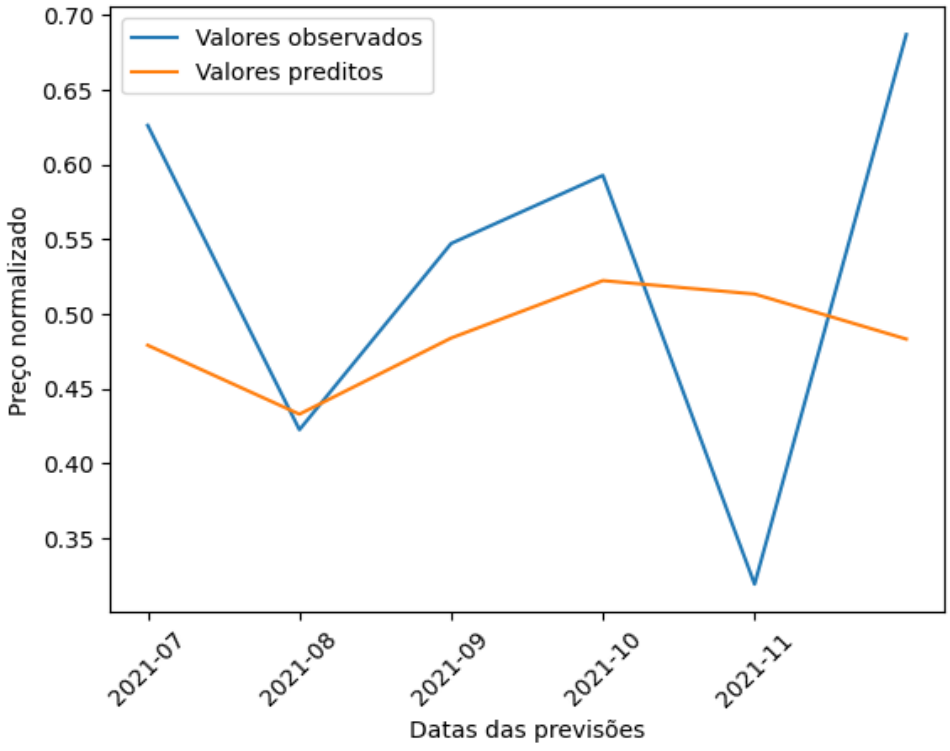
Os Gráficos 20 e 21 apresentam os gráficos com os valores previstos pelo modelo LSTM e os valores reais.

Gráfico 20 - Séries de valores reais e de teste para a previsão de preços no Ceagesp pelo método LSTM



Fonte: O autor

Gráfico 21 - Séries de valores reais e de teste para a previsão de preços no Ceasa/PR pelo método LSTM



Fonte: O autor

Dos gráficos é possível depreender que o modelo não foi eficaz em capturar os movimentos de alta e baixa da série. Guo et al. (2022) também identificaram o problema da qualidade de histerese em modelo LSTM. Assim, o método pode não ser uma boa escolha para análise de oscilações de curto prazo no conjunto de dados empregado neste estudo.

Para Greff et al. (2017), o método LSTM apresenta limitações, como a necessidade de ajuste de hiperparâmetros, que pode ser uma tarefa desafiadora em algumas aplicações, e a possibilidade de *overfitting*. Outra limitação é o tempo necessário para o treinamento em grandes conjuntos de dados.

5.4.6 CNN

Assim como o método LSTM, as CNNs não fornecem uma equação explícita para descrever a relação entre as variáveis de entrada e de saída. Em vez disso, as redes aprendem as relações entre os fatores de forma não linear, ajustando os pesos das camadas convolucionais e totalmente conectadas por meio de um processo de treinamento.

A fim de facilitar a interpretação dos parâmetros e a análise dos pesos do modelo, a Tabela 17 disponibiliza os valores de importância das variáveis e seus respectivos atrasos.

Tabela 17 - Importância das variáveis no modelo CNN para São Paulo e Curitiba

(continua)

Variável	Atraso	São Paulo	Curitiba
Índice de Atividade do Comércio	0	-0,0013	-0,0349
Índice de Atividade do Comércio	1	-0,0001	0,0102
Índice de Atividade do Comércio	2	0,0003	0,0026
Índice de Atividade do Comércio	3	0,0005	0,0041
IPCA	0	0,0019	0,0033
IPCA	1	0,0146	0,0006
IPCA	2	0,0077	-0,0118
IPCA	3	0,0006	-0,0014
Precipitação	0	-0,0060	-0,0004
Precipitação	1	-0,0043	0,0057
Precipitação	2	-0,0024	-0,0046
Precipitação	3	0,0079	-0,0039
Preço do tomate	1	-0,0027	0,0022

Tabela 17 - Importância das variáveis no modelo CNN para São Paulo e Curitiba

(conclusão)

Variável	Atraso	São Paulo	Curitiba
Preço do tomate	2	0,0368	0,0022
Preço do tomate	3	-0,0405	0,0026
Quantidade comercializada	0	-0,0282	-0,0065
Quantidade comercializada	1	-0,0025	-0,0151
Quantidade comercializada	2	-0,0016	0,0029
Quantidade comercializada	3	0,0000	0,0004
Temperatura	0	-0,0006	-0,0100
Temperatura	1	0,0048	-0,0084
Temperatura	2	-0,0116	-0,0026
Temperatura	3	0,0069	0,0043

Fonte: O autor

As variáveis atividade do comércio e quantidade comercializada, sem atrasos, têm um efeito negativo pronunciado, tanto em São Paulo quanto em Curitiba, logo, estão associadas a uma redução no desempenho do modelo quando permutadas.

Como ocorreu para o modelo LSTM, a variável preço com atraso de dois meses tem um efeito positivo significativo em São Paulo, por isso, esse fator está relacionado a um melhor desempenho do modelo na tarefa de previsão. A variável IPCA no mês anterior também possui efeito positivo para esse mercado.

Diferente da capital paulista, onde a atividade do comércio com um mês de atraso apresentou importância pouco significativa, para Curitiba, o valor de importância é 0,0102, dessa forma, a variável teve um impacto maior no desempenho do modelo para o Paraná. Na sequência, com importância significativamente menor, está a série de precipitação, com defasagem de um mês.

Os resultados das métricas de desempenho RMSE e MAPE, do modelo CNN para previsão de preços de tomate em São Paulo e Curitiba, são apresentados na Tabela 18.

Tabela 18 - Resultado das métricas de desempenho RMSE e MAPE do modelo CNN

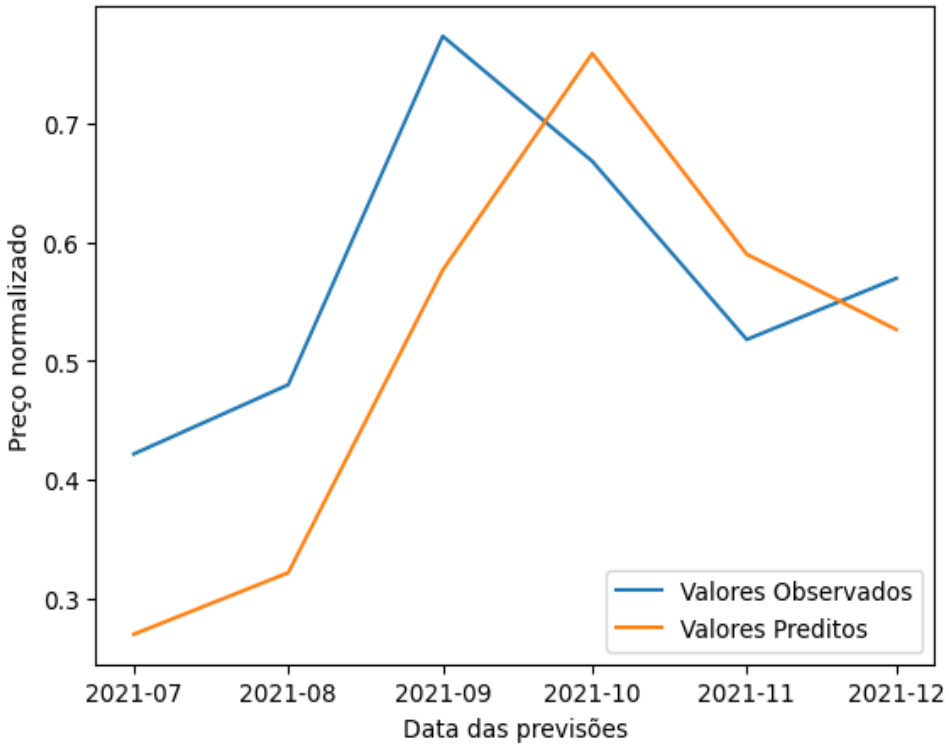
Local	RMSE	MAPE
São Paulo	0,21	0,22
Curitiba	0,16	0,13

Fonte: O autor

Os valores de MAPE mostram que modelo de redes convolucionais foi mais eficaz em prever valores para a série de preços de tomate em Curitiba, com desempenho razoável, do

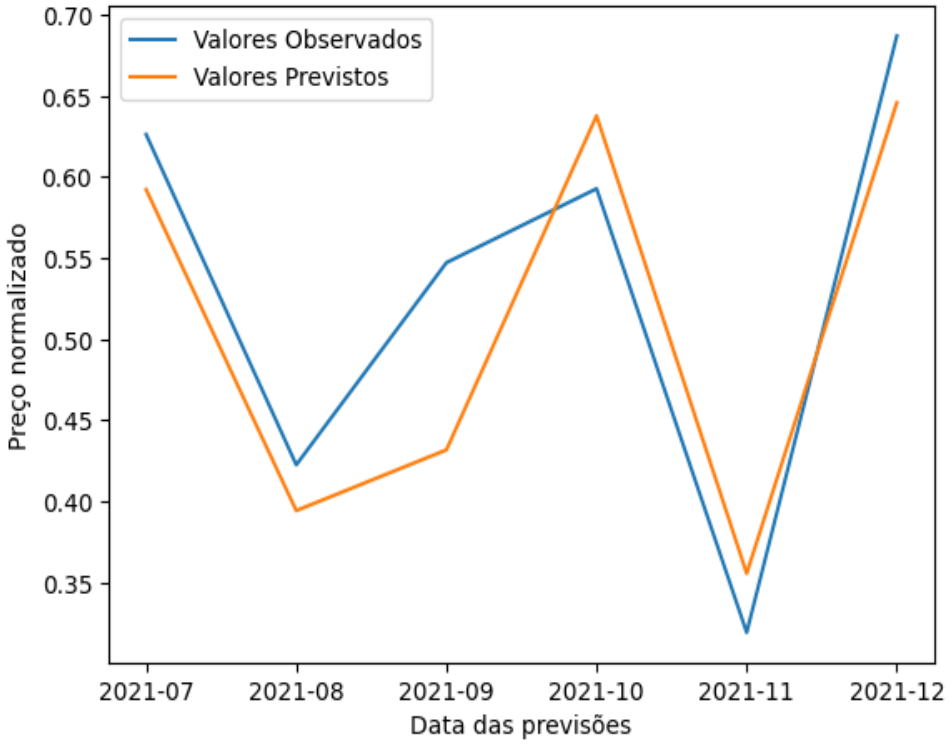
que em São Paulo, com baixa precisão. Essa mesma interpretação pode ser confirmada pela análise dos Gráficos 22 e 23, os quais evidenciam que o método CNN conseguiu captar com maior precisão as flutuações de preço na capital paranaense.

Gráfico 22 - Séries de valores reais e de teste para a previsão de preços no Ceagesp pelo método CNN



Fonte: O autor

Gráfico 23 - Séries de valores reais e de teste para a previsão de preços no Ceasa/PR pelo método CNN



Fonte: O autor

5.4.7 XGBoost

O XGBoost aplicado ao problema de regressão em tela é um método baseado em árvores de decisão e usa o algoritmo de *boosting* para melhorar o desempenho do modelo. Ele não segue um método estatístico explícito e, por isso, não tem uma equação de saída específica.

Com o objetivo de simplificar a compreensão dos parâmetros e facilitar a análise dos pesos do modelo, a Tabela 19 apresenta os valores de importância das variáveis com seus respectivos atrasos.

Tabela 19 - Importância das variáveis no modelo XGBoost para São Paulo e Curitiba

(continua)

Variável	Atraso	São Paulo	Curitiba
Índice de Atividade do Comércio	0	0,011	0,038
Índice de Atividade do Comércio	1	0,036	0,051
Índice de Atividade do Comércio	2	0,045	0,052
Índice de Atividade do Comércio	3	0,050	0,047

Tabela 19 - Importância das variáveis no modelo XGBoost para São Paulo e Curitiba

(conclusão)

Variável	Atraso	São Paulo	Curitiba
IPCA	0	0,032	0,014
IPCA	1	0,060	0,045
IPCA	2	0,037	0,026
IPCA	3	0,059	0,046
Precipitação	0	0,022	0,020
Precipitação	1	0,038	0,043
Precipitação	2	0,037	0,041
Precipitação	3	0,027	0,055
Preço do tomate	1	0,127	0,053
Preço do tomate	2	0,087	0,055
Preço do tomate	3	0,042	0,050
Quantidade comercializada	0	0,040	0,042
Quantidade comercializada	1	0,044	0,069
Quantidade comercializada	2	0,048	0,061
Quantidade comercializada	3	0,040	0,029
Temperatura	0	0,024	0,044
Temperatura	1	0,029	0,037
Temperatura	2	0,040	0,044
Temperatura	3	0,026	0,039

Fonte: O autor

Assim como para os demais métodos de aprendizado de máquina discutidos, em São Paulo as variáveis de preço do tomate em períodos anteriores são significativas para a tarefa de previsão com o modelo baseado em árvores de decisão. Outras variáveis importantes foram o índice de preço ao consumidor e atividade do comércio. A precipitação e temperatura, por outro lado, foram aquelas com menor relevância.

Para o Ceasa/PR a importância das variáveis para o foi equilibrada, o que indica boa capacidade do modelo em capturar as contribuições das séries. A quantidade comercializada e o preço observados há dois períodos foram as variáveis mais importantes nesse contexto.

A Tabela 20 exibe os resultados obtidos pelo modelo XGBoost na predição de preços de tomate no atacado em São Paulo e Curitiba, avaliados pelas métricas RMSE e MAPE.

Tabela 20 - Resultado das métricas de desempenho RMSE e MAPE do modelo XGBoost

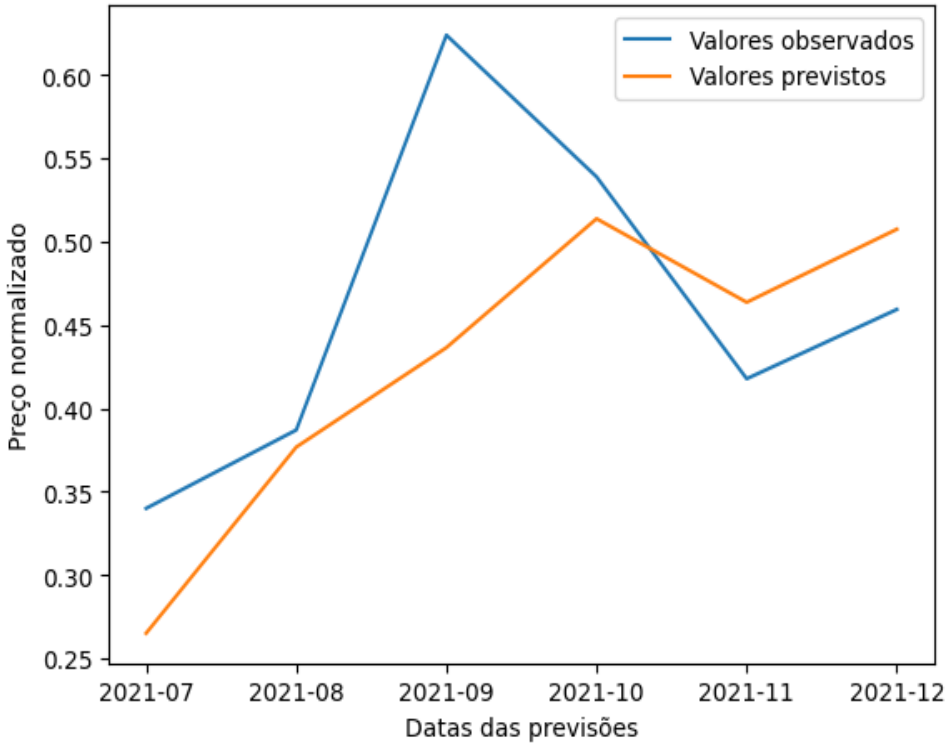
Local	RMSE	MAPE
São Paulo	0,09	0,13
Curitiba	0,11	0,20

Fonte: O autor

A partir dos resultados, conclui-se que o modelo apresentou desempenho superior ao prever séries para o Ceagesp. Para ambas as cidades os modelos podem ser considerados razoáveis.

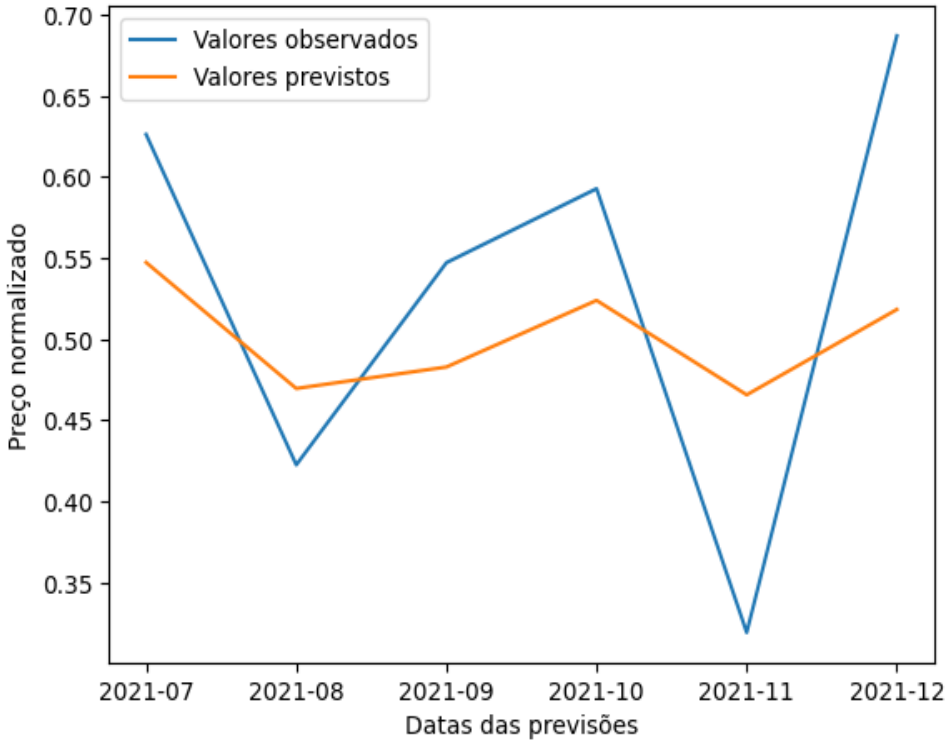
Os Gráficos 24 e 25 apresentados indicam que, apesar de registrarem valores razoáveis para o erro absoluto médio, os modelos não foram capazes de acompanhar a formação de picos e vales das séries de preços, o que pode levar a previsões imprecisas em momentos de alta volatilidade (HYNDMAN; ATHANASOPOULOS, 2018).

Gráfico 24 - Séries de valores reais e de teste para a previsão de preços no Ceagesp pelo método XGBoost



Fonte: O autor

Gráfico 25 - Séries de valores reais e de teste para a previsão de preços no Ceasa/PR pelo método XGBoost



Fonte: O autor

5.5 ANÁLISE DE RESÍDUOS

A Tabela 21 apresenta os resultados da estatística Q e valor-p do teste de Ljung-Box para determinar se há autocorrelação residual na série de valores previstos pelos modelos discutidos.

Tabela 21 - Resultados do teste de Ljung-Box para autocorrelação residual nas séries previstas

(continua)

		São Paulo		Curitiba	
Método	**Atraso**	**Estatística do teste**	**Valor-p**	**Estatística do teste**	**Valor-p**
ARIMA	1	0,03	0,86	0,02	0,88
	2	0,41	0,81	0,29	0,86
	3	1,25	0,74	0,32	0,96
	4	2,74	0,60	5,11	0,28
	5	3,09	0,69	6,04	0,30

Tabela 21 - Resultados do teste de Ljung-Box para autocorrelação residual nas séries previstas

(conclusão)

Método	Atraso	São Paulo		Curitiba	
		Estatística do teste	Valor-p	Estatística do teste	Valor-p
SARIMA	1	1,75	0,19	3,69	0,05
	2	4,52	0,10	3,69	0,16
	3	4,68	0,20	3,81	0,28
	4	5,72	0,22	5,11	0,28
	5	5,80	0,33	5,15	0,40
ARIMAX	1	0,24	0,63	0,00	0,95
	2	0,96	0,62	0,01	0,99
	3	1,21	0,75	0,41	0,94
	4	1,57	0,81	3,31	0,51
	5	1,87	0,87	3,87	0,57
SVR	1	0,46	0,50	1,09	0,30
	2	0,50	0,78	1,82	0,40
	3	4,39	0,22	2,46	0,48
	4	4,73	0,32	3,05	0,55
	5	4,88	0,43	3,13	0,68
LSTM	1	0,94	0,33	2,54	0,11
	2	3,50	0,17	2,56	0,28
	3	5,16	0,16	3,99	0,26
	4	5,18	0,27	8,00	0,09
	5	5,18	0,39	9,32	0,10
CNN	1	1,00	0,32	0,23	0,63
	2	1,47	0,48	1,27	0,53
	3	3,95	0,27	1,35	0,72
	4	4,92	0,30	2,29	0,68
	5	4,97	0,42	2,57	0,77

Fonte: O autor

O método XGBoost não pressupõe uma estrutura de autocorrelação nos resíduos, ao contrário dos demais. As árvores de decisão dividem o espaço de características em regiões retangulares e aprendem relações não lineares nos subespaços correspondentes. Como

resultado, os resíduos são tipicamente mais difíceis de avaliar quanto à autocorrelação (CHEN; GUESTRIN, 2016).

Após análise da Tabela é possível afirmar que, para todos os modelos, não há evidências suficientes para rejeitar a hipótese nula de que não há autocorrelação residual nos cinco atrasos possíveis para as séries de preços previstos para as centrais de abastecimento de São Paulo e Curitiba. Isso sugere que os modelos estão capturando adequadamente a estrutura de dependência nas séries temporais.

5.6 TESTE DE WILCOXON

A Tabela 22 compila os resultados das métricas de desempenho para os modelos utilizados neste estudo.

Tabela 22 - Resultados das métricas de desempenho, RMSE e MAPE, para os modelos candidatos

	RMSE		**MAPE**	
Método	**São Paulo**	**Curitiba**	**São Paulo**	**Curitiba**
ARIMA	0,22	0,14	0,43	0,24
SARIMA	0,26	0,16	0,51	0,30
ARIMAX	0,21	0,09	0,41	0,17
SVR	0,08	0,08	0,10	0,12
LSTM	0,12	0,12	0,23	0,22
CNN	0,21	0,16	0,22	0,13
XGBoost	0,09	0,11	0,13	0,20

Fonte: O autor

Os resultados do teste Wilcoxon sobre a significância das diferenças entre os valores reais e aqueles previstos por modelo, a 5% de probabilidade, são apresentados nas Tabelas 23 e 24. O valor-p indica a probabilidade de obter um resultado tão, ou, mais extremo que o observado, assumindo que não há diferença real.

Tabela 23 - Resultados de valor-p e avaliação de significância, a 5% de probabilidade, da diferença entre os erros de previsão dos modelos aplicados aos preços em São Paulo

Modelo 1	Modelo 2	Valor-p	Significância
SVR	XGBoost	0,0312	Sim
SVR	LSTM	0,0312	Sim
SVR	CNN	0,0312	Sim
SVR	ARIMAX	0,0312	Sim
SVR	ARIMA	0,0312	Sim
SVR	SARIMA	0,0312	Sim
XGBoost	LSTM	0,4375	Não
XGBoost	CNN	0.5625	Não
XGBoost	ARIMAX	0,0625	Não
XGBoost	ARIMA	0,0312	Sim
XGBoost	SARIMA	0,0312	Sim
LSTM	CNN	0,0625	Não
LSTM	ARIMAX	0,0625	Não
LSTM	ARIMA	0,0312	Sim
LSTM	SARIMA	0,0312	Sim
CNN	ARIMAX	0,0625	Não
CNN	ARIMA	0,0312	Sim
CNN	SARIMA	0,0312	Sim
ARIMAX	ARIMA	0,8438	Não
ARIMAX	SARIMA	0,0312	Sim
ARIMA	SARIMA	0,0312	Sim

Fonte: O autor

Tabela 24 - Resultados de valor-p e avaliação de significância, a 5% de probabilidade, da diferença entre os erros de previsão dos modelos aplicados aos preços em Curitiba

(continua)

Modelo 1	Modelo 2	Valor-p	Significância
SVR	XGBoost	0,6875	Não
SVR	CNN	0,2188	Não
SVR	LSTM	0,3125	Não
SVR	ARIMAX	0,8438	Não
SVR	ARIMA	0,3125	Não
SVR	SARIMA	0,4375	Não
XGBoost	CNN	0,4375	Não
XGBoost	LSTM	0,0938	Não
XGBoost	ARIMAX	0,6875	Não

Tabela 24 - Resultados de valor-p e avaliação de significância, a 5% de probabilidade, da diferença entre os erros de previsão dos modelos aplicados aos preços em Curitiba

(conclusão)

Modelo 1	Modelo 2	Valor-p	Significância
XGBoost	ARIMA	0,2188	Não
XGBoost	SARIMA	0,2188	Não
CNN	LSTM	0,1562	Não
CNN	ARIMAX	0,1562	Não
CNN	ARIMA	0,1562	Não
CNN	SARIMA	0,2188	Não
LSTM	ARIMAX	0,1562	Não
LSTM	ARIMA	0,5625	Não
LSTM	SARIMA	0,5625	Não
ARIMAX	ARIMA	0,4375	Não
ARIMAX	SARIMA	0,3125	Não
ARIMA	SARIMA	0,5625	Não

Fonte: O autor

Com base nos erros de previsão de preços de tomate em São Paulo, observa-se que o método SVR obteve desempenho significativamente superior aos demais. Isso indica que as diferenças observadas nas métricas de desempenho não foram devidas ao acaso. Esse resultado está em linha com aquele obtido para RMSE e MAPE, que indicam o método de vetor de suporte com menores erros.

Para Curitiba, o teste de Wilcoxon não encontrou diferenças estatisticamente significativas nos *rankings* dos erros dos modelos. Nesse caso, é importante considerar a amplitude das diferenças nos valores de RMSE e MAPE apresentados na Tabela 22. A análise das métricas aponta que o modelo de saída do SVR é superior. Portanto, mesmo que o teste estatístico não encontre diferenças, o método de vetores de suporte é o mais preciso, assim como para o conjunto de dados da capital paulista.

A superioridade do *support vector regression* pode ser explicada pela capacidade do método de lidar com conjuntos de dados complexos, não lineares e com alta dimensionalidade (SUYKENS; VANDEWALLE, 1999). Além disso, o fato de o SVR ter superado os métodos tradicionais de séries temporais, como ARIMA e SARIMA, sugere que ele é capaz de capturar padrões mais complexos e sutis nos dados.

O resultado ora obtido não confirma aquele apresentado por Yuan e Ling (2020). Nesse estudo, entre os métodos SVR, LSTM, ARIMA e XGBoost, o modelo de vetor de suporte foi o menos preciso para prever preços de tomate, quando avaliado pela métrica erro quadrático médio. No trabalho de Purohit et al. (2021) o modelo com vetores de suporte foi

menos preciso do que o ARIMA. Contrariando esse último estudo, os resultados de Paul et al. (2022) mostram que o modelo autorregressivo foi superado pelo SVR.

Para as séries previstas com dados do Ceagesp, os métodos multivariados foram consistentemente mais precisos do que ARIMA e SARIMA. Nas previsões para o Ceasa/PR, as métricas de desempenho indicam o mesmo resultado. Assim, o método ARIMAX é exceção entre os clássicos, com resultados comparáveis a modelos de aprendizado de máquina.

A escolha do método mais adequado deve levar em consideração não apenas a precisão alcançada, mas também outros fatores, como a disponibilidade de recursos e a facilidade de interpretação dos resultados.

Para qualquer modelo multivariado candidato neste estudo, o desempenho pode ser afetado pela escolha das variáveis explicativas e pelos parâmetros de ajuste, o que sugere a necessidade de uma análise mais aprofundada e experimentação cuidadosa. Além disso, é importante destacar que a previsão de preços é uma tarefa complexa e difícil de ser realizada com alta precisão, dadas as diversas fontes de incerteza e volatilidade presentes no mercado de produtos agrícolas.

6 CONCLUSÕES

A formação de preços agrícolas é influenciada por diversos fatores, tais como variáveis econômicas e climáticas.

A previsão de preços de tomate pode auxiliar a tomada de decisão de agentes envolvidos, desde a produção até o consumidor final.

A comparação dos resultados entre os diferentes métodos clássicos e de aprendizado de máquina na previsão de preços de tomate no atacado mostrou que, o modelo gerado pelo SVR apresentou o melhor desempenho em termos de precisão. Entre os métodos clássicos, o ARIMAX foi aquele com melhor desempenho.

As variáveis mais relevantes para o modelo de vetor de suporte foram os preços anteriores da variável alvo, a quantidade de tomate comercializada e a temperatura.

A escolha do método mais adequado pode variar de acordo com o objetivo da previsão, o período de tempo previsto, a disponibilidade e qualidade dos dados, entre outros fatores. Portanto, é necessário avaliar cuidadosamente o desempenho e considerar as limitações e suposições de cada um antes de selecionar um modelo para prever os preços futuros de tomate.

Estudos futuros podem explorar o uso de métodos híbridos e de dados climáticos de regiões produtoras para melhorar a precisão da previsão de preços de tomate e de outros produtos agrícolas. Outra indicação é a análise da influência da correção pelo índice de inflação nos resultados obtidos. Essa análise permitirá uma compreensão mais profunda de como a inflação afeta as previsões de preços e se a correção é ou não benéfica em termos práticos.

REFERÊNCIAS

ADANACIOGLU, H.; YERCAN, M. An analysis of tomato prices at wholesale level in Turkey: an application of SARIMA model. **Custos e@gronegócio on line**, v. 8, n. 4, p. 52-75, 2012.

ANAYA, M. **Por que o real se desvalorizou mais se covid-19 também afeta outros países?** UOL Economia, São Paulo, 22 mar. 2020. Disponível em: https://economia.uol.com.br/cotacoes/noticias/redacao/2020/03/22/real-dolar-coronavirus-queda.htm. Acesso em: 20 de junho de 2023.

ANGGRAENI, W.; MAHANANTO, F.; SARI, A. Q.; ZAINI, Z.; ANDRI, K. B. Forecasting the price of Indonesia's rice using hybrid artificial neural network and autoregressive integrated moving average (Hybrid NNs-ARIMAX) with exogenous variables. **Procedia Computer Science**, v. 161, p. 677-686, 2019.

BALTAR, B. P. Análise temporal dos preços da commodity cobre usando o Modelo Box e Jenkins. Dissertação (Mestrado em Engenharia de Produção) - **Pontifícia Universidade Católica do Rio de Janeiro**, Rio de Janeiro, 2009.

BERA, A. K.; JARQUE, C. M. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. **Economics Letters.** 1980.

BLANCHARD, O. J.; SHEEN, J. R. Macroeconomics: Australian Edition. 1. ed. **Pearson,** Australia, 2012.

BOX, G. E. P.; JENKINS, G. M. Time Series Analysis: Forecasting and Control. **Wiley**, 1976.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. Time Series Analysis: Forecasting and Control. **John Wiley; Sons**, 2015.

BRASIL. Ministério da Agricultura e do Abastecimento. **Brazil - Agricultural Policies. Brasília**, 2008. Disponível em: https://www.gov.br/agricultura/pt-br/assuntos/politica-agricola/outras-publicacoes/brazil-agricultural-policies-2008.pdf. Acesso em: 05 de junho de 2023.

BRAZ, A. Itens de peso, os vilões da inflação. **Conjuntura Econômica**. p. 113-114, 2005.

BROCKWELL, P. J.; DAVIS, R. A. Introduction to Time Series and Forecasting. **Springer**, 2016.

CHATFIELD, C. The Analysis of Time Series: An Introduction, Sixth Edition. **Chapman and Hall/CRC**, 2003.

CHAUHAN B. S., KAUR P., MAHAJAN G., RANDHAWA R. K., SINGH H., KANG M. S. **Global warming and its possible impact on agriculture in India**. p. 65–121, 2014.

CHEN, P.; YE, H. Short-Term Forecast of Agricultural Prices Using CNN+LSTM. 7th International Conference on Intelligent Information Processing. Bucharest, Romania: **Association for Computing Machinery**, 2022.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. **International Conference on Knowledge Discovery and Data Mining**. 2016.

COMPANHIA NACIONAL DE ABASTECIMENTO. CONAB. **Perspectivas para a agropecuária.** Vol. 9. Brasília, 2021.

COMPANHIA NACIONAL DE ABASTECIMENTO. CONAB. Tomate: Análise dos Indicadores da Produção e Comercialização no Mercado Mundial, Brasileiro e Catarinense. **Conab**. V.21. Brasília, 2019.

CONOVER, W. J. Practical Nonparametric Statistics. New York. **John Wiley & Sons**, 1999.

CRYER, J. D.; CHAN, K. S. Time Series Analysis: With Applications in R. **Springer**, 2008.

DAREKAR, A.; REDDY, A. Predicting market price of soybean in major India studies through ARIMA model. **Journal of Food Legumes**, Hyderabad, v. 30, n. 2, p. 73-76, 2017.

DAVENPORT, F.; FUNK, C. Using time series structural characteristics to analyze grain prices in food insecure countries. **Food Security**, v. 7, p. 1055-1070, 2015.

DEVI, M.; KUMAR, J.; MALIK, D. P.; MISHRA, P. Forecasting of wheat production in Haryana using hybrid time series model. **Journal of Agriculture and Food Research**, v. 5, p. 1-7, 2021.

DICKEY, D. A.; FULLER, W. A. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. **Journal of the American Statistical Association**, v. 74, n. 366, p. 427-431, 1979.

DILLON, W. R.; GOLDSTEIN, M. Multivariate Analysis. 2. ed. New York. **John Wiley & Sons**, 1984.

DRUCKER, H.; BURGES, C. J. C.; KAUFMAN, L.; SMOLA, A. J.; VAPNIK, V. Support Vector Regression Machines. **Advances in Neural Information Processing Systems**, v. 9, 1997.

FAHAD, S.; BAJWA, A. A.; NAZIR, U.; ANJUM, S. A.; FAROOQ, A.; ZOHAIB, A.; SADIA, S.; NASIM, W.; ADKINS, S.; SAUD, S.; IHSAN, M. Z.; ALHARBY, H.; WU, C.; WANG, D.; HUANG, J. Crop Production under Drought and Heat Stress: Plant Responses and Management Options. **Frontiers in Plant Science**, v. 8, p. 1147, 2017.

FAO. Organização das Nações Unidas para Alimentação e Agricultura. The State of Food and Agriculture. **Food Systems for Better Nutrition.** Roma, 2013.

GALLO, GUSTAVO. Análise da sazonalidade do preço do tomate no Ceasa da Grande Florianópolis. **Universidade Federal de Santa Catarina**. Florianópolis, 2007.

GÉRON, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. **O'Reilly Media**, 2019.

GERS, F. A.; SCHMIDHUBER, J.; CUMMINS, F. Learning to forget: Continual prediction with LSTM. **Neural Computation**, v. 12, n. 10, p. 2451-2471, 2000.

Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning. **MIT Press**, 2016.

GREFF, K. et al. LSTM: A search space odyssey. **IEEE Transactions on Neural Networks and Learning Systems**, v. 28, n. 10, p. 2222-2232, Oct. 2017.

GUO, Y.; TANG, D.; TANG, W.; YANG, S.; TANG, Q.; FENG, Y.; ZHANG, F. Agricultural Price Prediction Based on Combined Forecasting Model under Spatial-Temporal Influencing Factors. **Sustainability**, v. 14, n. 17, p. 10483, 2022.

HAMULCZUK, M.; GRUDKOWSKA, S.; GĘDEK, S.; KLIMKOWSKI, C.; STANKO, S. Essential econometric methods of forecasting agricultural commodity prices. **Institute of Agricultural and Food Economics of Poland.** 2013.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Nova York. **Springer**, 2009.

HAYKIN, S. Neural Networks and Learning Machines. **Prentice Hall**, 2009.

HIGGINS, J. P. T.; GREEN, S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. **The Cochrane Collaboration**, 2011.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735-1780, 1997.

HYNDMAN, R. J.; ATHANASOPOULOS, G. Forecasting: principles and practice. **OTexts**, 2018.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **PAM - Produção Agrícola Municipal**. Rio de Janeiro, 2021. Disponível em: https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/9103-producao-agricola-municipal.html. Acesso em: 02 maio 2023.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Pesquisa de Orçamentos Familiares 2017-2018: análise da disponibilidade domiciliar de alimentos e do estado nutricional no Brasil. Rio de Janeiro: **IBGE**, p. 168, 2020.

JIN, D.; YIN, H.; GU, Y.; YOO, S. J. Forecasting of vegetable prices using STL-LSTM method. **6th International Conference on Systems and Informatics**, p. 866-871, 2019.

JOHNSON, R. A.; WICHERN, D. W. Applied Multivariate Statistical Analysis. 6. ed. **Pearson**, 2014.

KAUR, L.; MISHRA, A. A comparative analysis of evolutionary algorithms for the prediction of software change. **INTERNATIONAL CONFERENCE ON INNOVATIONS IN INFORMATION TECHNOLOGY (IIT)**, 2018.

KURUMATANI, K. Time series forecasting of agricultural product prices based on recurrent neural networks and its evaluation method. **Appl. Sci**. 2, 1434, 2020.

LAI, G.; XIE, X.; LIU, H.; CHEN, J. Modeling long- and short-term temporal patterns with deep neural networks. **IEEE Transactions on Neural Networks and Learning Systems**, v. 29, n. 10, p. 5017-5029, 2018.

LIU, D.; TANG, Z.; CAI, Y. A Hybrid Model for China's Soybean Spot Price Prediction by Integrating CEEMDAN with Fuzzy Entropy Clustering and CNN-GRU-Attention. **Sustainability**, v. 14, p. 15522, 2022.

LUNDBERG, S. M.; LEE, S. I. A unified approach to interpreting model predictions. Advances in neural information processing systems. **CA: Neural Information Processing Systems Foundation**, p. 4765-4774, 2017

LY, R.; TRAORE, F.; DIA, K. Forecasting commodity prices using long-short-term memory neural networks. Vol. 2000. **Intl Food Policy Res**. Washington, DC:, 2021.

MAGALHÃES, M. M. Precedência temporal dos preços de commodities agrícolas em relação ao nível de atividade econômica: uma abordagem por meio de modelos VAR. Tese de Doutorado, **Universidade de São Paulo**. 2011.

MAKRIDAKIS, S.; WHEELWRIGHT, S. C.; HYNDMAN, R. J. Forecasting methods and applications. 3. ed. **New York: John Wiley**, 1998.

MALDONADO, Sebastián; GONZALEZ, Agustin; CRONE, Sven. Automatic time series analysis for electric load forecasting via support vector regression**. Applied Soft Computing**, v. 83, p. 105616, 2019.

MCKINNEY, WES. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. **O'Reilly Media**, 2017.

MILLY, P. C. D.; BETANCOURT, J.; FALKENMARK, M.; HIRSCH, R. M.; KUNDZEWICZ, Z. W.; LETTENMAIER, D. P.; STOUFFER, R. J. Stationarity Is Dead: Whither Water Management? **Science**, v. 319, n. 5863, p. 573-574. 2008.

MONTGOMERY, D. C.; JOHNSON, L. A.; GARDINER, J. S. Forecasting and Time Series Analysis. **McGraw-Hill Education**, 2015.

MORETTIN, P. A.; TOLOI, C. M. C. Análise de Séries Temporais. 2. ed. São Paulo: **Blucher**, 2018.

MURPHY, K. P. Machine learning: a probabilistic perspective. **The MIT Press**, 2012.

MURUGESAN, R.; MISHRA, E.; KRISHNAN, A. H. Forecasting agricultural commodities prices using deep learning-based models: basic LSTM, bi-LSTM, stacked LSTM, CNN LSTM, and convolutional LSTM. I**nternational Journal of Sustainable Agricultural Management and Informatics**, v. 8, n. 3, 2022.

NAZLIOGLU, S.; SOYTAS, U. World oil prices and agricultural commodity prices: Evidence from an emerging market. **Energy Economics**, v. 33, n. 3, p. 488-496, 2011.

PANDAS. **About pandas**. Disponível em: https://pandas.pydata.org/about/. Acesso em: 24 jun. 2023.

PAUL, R. K.; YEASIN, M.; KUMAR, P.; KUMAR, P.; BALASUBRAMANIAN, M. Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. **PLOS ONE**, 17(7), e0270553, 2022.

PÉRA, T.G.; COSTA, E.L.; CAIXETA-FILHO, J.V. Impactos dos reajustes dos preços de óleo diesel na logística do agronegócio brasileiro no período de janeiro/2017 a maio/2018. **Logística do Agronegócio Desafios e Oportunidades**, 2018.

PÉREZ, A. G. Introduction to Time Series Analysis: An Applied Guide in R. **Wiley**, 2021.

PUROHIT, S.K.; PANIGRAHI, S.; SETHY, P.K.; BEHERA, S.K. Time Series Forecasting of Price of Agricultural Products Using Hybrid Methods. **Applied Artificial Intelligence**, 35(15), p. 1388-1406, 2021.

RAO, J. M. Agricultural Supply Response: A Survey. **Agricultural Economics**, v. 3, p. 1-22, 1989.

REDDY, A. A. Price Forecasting of Tomatoes. **International Journal of Vegetable Science**, 25(2), p. 176-184, 2019.

REN, C.; LUO, X. Comparing the Performance of ARIMA and Exponential Smoothing Models in Forecasting Grain Prices in China. **Journal of Agricultural and Applied Economics**, v. 50, n. 1, p. 65-80, 2018.

REZENDE A. L. B.; FILHO C. B. A.; MARTINS G. E. I. M.; COSTA C. C.; FELTRIM L. A. **V**iabilidade econômica das culturas de pimentão,repolho, alface, rabanete e rúcula em cultivo consorciado, na primavera-verão**.** Jabotical, estado de São Paulo. **Informações Econômicas,** p. 22-36, 2005.

REZENDE, G. C. de; SILVA, E. A. da; SILVA, P. C. G. da; SILVA, R. L. da. O efeito dos preços dos insumos na formação de preços agrícolas no Brasil. **Revista de Política Agrícola**, Brasília, v. 24, n. 4, p. 5-22, 2015.

ROSNER, B. Fundamentals of Biostatistics. T**homson Brooks/Cole**. Belmont, CA., 2006.

SAID, S. E.; DICKEY, D. A. Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. **Biometrika**, v. 71, n. 3, p. 599-607, 1984.

SANTOS, G. C.; BARBOZA, F.; VEIGA, A. C. P.; SILVA, M. F. Forecasting Brazilian Ethanol Spot Prices Using LSTM. **Energies**, v. 14, p. 7987, 2021.

SCIKIT-LEARN. permutation_importance. **Scikit-learn: Machine Learning in Python**. https://scikit-learn.org/stable/modules/generated/sklearn.inspection.permutation_importance.html, 2021. Acessado em: 07 de maio de 2023.

SERRA, T.; GIL, J. M. Forecasting agricultural commodity prices with Asymmetric Error Component Autoregressive models. **Applied Economics**, v. 39, n. 9, p. 1183-1193, 2007.

SHENGWEI, W.; YANNI, L.; JIAYU, Z.; JIAJIA, L. Agricultural price fluctuation model based on SVR. **INTERNATIONAL CONFERENCE ON MODELLING, IDENTIFICATION AND CONTROL**, p. 545-550, 2017.

SHUMWAY, R. H.; STOFFER, D. S. Time Series Analysis and Its Applications: With R Examples. **Springer Texts in Statistics**, Springer, 2017.

SIEGEL, S.;CASTELLAN, N. J. Nonparametric statistics for the behavioral sciences. **New York: McGraw-Hill**, 1988.

SILVA, A.; SANTOS, B.; PEREIRA, C. Ano climático de referência para Curitiba: comparação entre dados de duas estações. **XV Encontro Nacional de Conforto no Ambiente Construído**, p. 123-456, 2019.

SINGH, S.; KAUR, H. Time series analysis of agricultural product prices: A review. **International Journal of Agricultural and Statistical Sciences**, v. 13, n. 2, p. 543-550, 2017.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, v. 14, n. 3, p. 199-222, 2004.

SOLOMON, M. R. Comportamento do Consumidor: Comprando, Tendo e Sendo**.** 12ª ed. **Pearson Education**. São Paulo, 2022.

SOUZA, R. S.; VIANA, J. G. A. Tendência histórica de preços pagos ao produtor na agricultura de grãos do Rio Grande do Sul, Brasil. **Cienc. Rural**, v. 37, n. 4, p. 1128-1133, 2007.

STOCK H. J. Time series: Economic forecasting**. International Encyclopedia of the Social &Behavioral Sciences**. 2001. Pág. 15721-15724.

SUYKENS, J. A. K.; VANDEWALLE, J. Least squares support vector machine classifiers. **Neural Processing Letters**, v. 9, n. 3, p. 293-300, 1999.

VANDERPLAS, JAKE. Python Data Science Handbook: Essential Tools for Working with Data. **O'Reilly Media**, 2016.

VAPNIK, V. The Nature of Statistical Learning Theory. **Springer**, 1995.

VEGA-MÁRQUEZ, B.; RUBIO-ESCUDERO, C.; NEPOMUCENO-CHAMORRO, I. A.; ARCOS-VARGAS, Á. Use of Deep Learning Architectures for Day-Ahead Electricity Price Forecasting over Different Time Periods in the Spanish Electricity Market. **Applied Sciences**, v. 11, n. 13, p. 6097, 2021.

VÉLEZ, C.;.; SALAZAR, J. D. Forecasting coffee prices in Colombia: a comparison of ARIMA and exponential smoothing models. **Applied Economics Letters**, v. 22, n. 3, p. 224-229, 2015.

WENG, Q.; LIU, R.; TAO, Z. Forecasting Tesla's Stock Price Using the ARIMA Model. **Proceedings of Business and Economic Studies**, v. 5, n. 5, 2022. ISSN Online: 2209-265X. ISSN Print: 2209-2641. Disponível em: http://ojs.bbwpublisher.com/index.php/PBES. Acesso em: 11 mar. 2023.

WITTEN, I. H.; FRANK, E.; HALL, M. A. Data Mining: Practical Machine Learning Tools and Techniques. **3. ed. San Francisco: Morgan Kaufmann**, 2016.

XIAO, W.; XU, C.; LIU, H.; LIU, X. A Hybrid LSTM-Based Ensemble Learning Approach for China Coastal Bulk Coal Freight Index Prediction. J**ournal of Advanced Transportation**, v. 2021, p. 1-23, 2021.

XIE, W.; YU, L.; XU, S.; WANG, S. A New Method for Crude Oil Price Forecasting Based on Support Vector Machines. **Lecture Notes in Computer Science**, vol 3994. Springer, Berlin, Heidelberg, p. 588-595. 2006.

YOO, T.; OH, I. Time series forecasting of agricultural products sales volumes based on seasonal long short-term memory. **Applied Sciences**, v. 10, n. 22, p. 8169, 2020.

YUAN, C. Z.; LING, S. K. Long short-term memory model based agriculture commodity price prediction application. **2nd International Conference on Information Technology and Computer Communications**. p. 43-49. 2020.

ZHANG, H.; YU, C.; LU, J.; GONG, P. Feature Selection Using Principal Component Analysis. **Journal of Chemometrics**, v. 23, n. 3-4, p. 125-136, 2009.