LVI SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
DE 04 ATÉ 07 NOVEMBRO 2024
FORTALEZA | CEARÁ

Comitê de Máquinas Heterogêneas e Redes Transformers na Previsão de Preços de Commodities: Evidências Empíricas

Henri Makika*

Romis Ribeiro de Faissol Attux*

João Marcos Travassos Romano*
(*) Faculdade de Engenharia Elétrica e de Computação – FEEC - Univ. Estadual de Campinas
Av. Albert Einstein, Nº 400 - Cidade Universitária, Campinas - SP, 13083-852, Brasil
h211042@dac.unicamp.br; attux@unicamp.br; jmromano@unicamp.br

RESUMO

A avaliação da precisão dos métodos de aprendizado de máquina na previsão de preços de commodities agrícolas desempenha um papel importante na pesquisa acadêmica. Desenvolver modelos robustos é fundamental, e a combinação de Comitês de Máquinas e Redes Transformers emerge como uma abordagem promissora. Este estudo visa, empiricamente, comparar a capacidade preditiva de Comitês de Máquinas Heterogêneas e modelos Transformers com outras abordagens, como DCNN, Bagging e Boosting, especialmente na previsão de preços da soja no contexto do agronegócio. Os resultados mostram que o Comitê-LSTM tem o melhor desempenho na previsão de curto prazo, enquanto a rede Transformer mostra um declínio gradual em seu desempenho ao longo do tempo, medido pelo NSE, em comparação com os outros modelos.

PALAVRAS CHAVE. Comitê Heterogêneo, Redes Transformers, Previsão de Séries Temporais, Preços da Soja.

Tópicos (IC - Inteligência Computacional, GF – Gestão Financeira)

ABSTRACT

The evaluation of the accuracy of machine learning methods in forecasting agricultural commodity prices is a research topic of great practical importance. Developing robust models is essential, and the combination of Machine Committees and Transformer Networks emerges as a promising approach. This study aims to empirically compare the predictive ability of Heterogeneous Machine Committees and Transformer models with other approaches, such as DCNN, Bagging, and Boosting, especially in forecasting soybean prices in the context of agribusiness. Results show that the Committee-LSTM performs best in short-term forecasting, while the Transformer network exhibits a gradual decline in performance over time, measured by NSE, compared to other models.

KEYWORDS. Heterogeneous Committee. Transformers. Time Series Forecasting. Soybean Prices.

Paper topics (Computational Intelligence, Financial Management)

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

1. Introdução

A cadeia de valor da soja desempenha um papel essencial na agricultura global, sendo uma das principais cadeias agrícolas do mundo. Na safra global de 2021/2022, a produção alcançou 355.588 milhões de toneladas em uma área plantada de 130.935 milhões de hectares [FAO, 2024], o que evidencia a sua importância. A demanda por soja, tanto no contexto mundial quanto no mercado interno brasileiro, é significativa. O Brasil, classificado como o segundo maior produtor mundial e líder em exportação, registrou uma produção de 131,00 milhões de toneladas na safra 2021/2022, em uma área plantada de 40,3 milhões de hectares [CONAB, 2024].

De acordo com a Companhia Nacional de Abastecimento (CONAB), o Brasil alcançou uma produção recorde de 154,6 milhões de toneladas de soja na safra 2022/2023, com uma área de 43,3 milhões de hectares. Esse aumento de 27,3 milhões de toneladas em relação à safra anterior consolida ainda mais o Brasil como o principal produtor e exportador de soja. Globalmente, 60% da produção de soja é destinada à exportação, enquanto os restantes 40% são consumidos internamente no Brasil. Nesse contexto, a previsão da série histórica de preços da soja assume uma importância capital no panorama econômico nacional. A precisão dessas previsões é fundamental para antecipar resultados e está intimamente ligada à exatidão do modelo utilizado. Sua relevância transcende várias áreas do conhecimento, já que as previsões são elementos essenciais para embasar a formulação de políticas públicas.

Com o intuito de aperfeiçoar a precisão dos modelos de regressão no contexto de previsão de séries temporais, diversas abordagens têm sido exploradas na literatura, sendo os comitês de regressores uma delas [Liu et al., 2020], [Zhou, 2012]. O princípio subjacente baseia-se nas ideias de diversificação e de ”dividir para conquistar”, superando as limitações dos modelos que operam de forma independente. Os comitês de modelos consistem em vários modelos básicos (sejam eles homogêneos ou heterogêneos) desenvolvidos para resolver o mesmo problema. Cada um desses modelos aprende características dos dados e gera previsões individuais, culminando em uma previsão final que é uma combinação ponderada das previsões individuais. Essa abordagem busca criar um sistema robusto que integra as previsões de todos os modelos básicos. Em geral, essa abordagem tem demonstrado resultados superiores em comparação com modelos de previsão individuais, pois exibe uma maior capacidade de generalização e adaptação a diferentes cenários [Divina et al., 2018], [Ribeiro e dos Santos Coelho, 2020].

Por outro lado, a arquitetura Transformer [Vaswani et al., 2017] é uma abordagem de aprendizagem profunda baseada em mecanismos de atenção. O mecanismo de atenção busca conferir aos modelos a capacidade de focar nos elementos mais relevantes de sequências extensas [Zeng et al., 2023], e tem demonstrado resultados interessantes na previsão de séries temporais.

Na pesquisa acadêmica, diversos estudos, como os referenciados em [Ribeiro e dos Santos Coelho, 2020], [Divina et al., 2018], [Zhou et al., 2021], investigaram a previsão dos preços de commodities agrícolas por meio da análise de dados históricos. Essas pesquisas apresentam similaridades em suas abordagens e estratégias de análise.

Em [Ribeiro e dos Santos Coelho, 2020], realiza-se uma comparação entre diversos modelos, incluindo Florestas Aleatórias (RF), Máquina de Reforço de Gradiente (GBM), Máquina de Reforço de Gradiente Extremo (XGB), Empilhamento (STACK), Máquina de Vetores de Suporte para Regressão (SVR), a Rede Neural Perceptron Multicamadas (MLP) e os K Vizinhos Mais Próximos (KNN). Esta análise é conduzida para a previsão de séries históricas mensais, especificamente os preços pagos aos produtores do estado do Paraná por uma saca de 60 kg de soja (Estudo de Caso 1) e trigo (Estudo de Caso 2). Os resultados indicam que o Erro Percentual Absoluto Médio (MAPE) foi inferior a 1% para os modelos XGB, STACK, KNN, SVR e RF nos Estudos de Caso 1 e 2. Esse desempenho destaca a eficácia desses modelos na previsão dos preços mencionados.

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

Em [Divina et al., 2018], realizou-se uma comparação entre um comitê de máquinas empilhado com modelos ANN, GBM, ARIMA e RF. As simulações foram executadas sobre um conjunto de dados abrangendo mais de nove anos, relatando o consumo de energia na Espanha. Os resultados revelaram que o comitê atingiu previsões altamente precisas, indicando assim que essa abordagem é apropriada para enfrentar o desafio de prever a carga de curto prazo. [Zhou et al., 2021] apresenta uma nova arquitetura de Transformer denominada Informer, que foi comparada com modelos como Transformer Vanilla, ARIMA, Prophet, DeepAR e LSTM. O estudo empregou um conjunto de diversas séries temporais provenientes de fontes variadas. Os resultados destacaram que o desempenho do Informer superou os demais modelos, especialmente em previsões de janelas temporais mais extensas.

O presente estudo tem como objetivo investigar a capacidade preditiva de comitês de máquinas heterogêneas, bem como da rede Transformer, por meio de uma comparação entre si. Ademais, são examinadas individualmente abordagens como Bagging, Boosting e redes neurais convolucionais Dilatadas (DCNN). As simulações são conduzidas utilizando dados históricos de preços diários da soja pagos aos produtores por sacas de 60 kg. Este experimento abrange um horizonte temporal de 30 dias à frente, empregando a abordagem direta de previsão. A escolha desse horizonte de previsão na área do agronegócio é motivada pelo fato de que a previsão de curto prazo permite que produtores de diversos portes desenvolvam estratégias de planejamento para atender a suas necessidades imediatas. Diversas métricas, como MAPE(%), MSE, RMSE e o Coeficiente de Eficiência de Nash-Sutcliffe (NSE), são utilizadas na avaliação dos métodos.

2. Modelos Teóricos

O problema de regressão não paramétrica pode ser descrito da seguinte maneira. Dado um conjunto de N amostras, representadas por pares $\boldsymbol{X_t}, \boldsymbol{y}_{t=1}^{T}$, onde cada $x_t \in \boldsymbol{X}$ é uma covariável e $y_t \in \mathbb{R}$ é uma resposta real, assumindo que as amostras são independentes e identicamente distribuídas (i.i.d.) de uma distribuição conjunta desconhecida $\mathbb{P}$ sobre $\boldsymbol{X} \times \mathbb{R}$. O objetivo é estimar uma função $\hat{f} : \boldsymbol{X} \to \mathbb{R}$, que pode ser utilizada para prever respostas futuras com base apenas nas covariáveis observadas. A qualidade da estimativa $\hat{f}$ é tipicamente avaliada em termos do erro quadrático médio de previsão, $J\varepsilon(\boldsymbol{w}) = \mathbb{E}[(Y - \hat{f}(X))^2]$, onde a expectativa condicional $f^*(x) = \mathbb{E}[Y|X = x]$ é o ponto ótimo. A literatura oferece uma variedade de estimadores [Hastie et al., 2009] associados à solução ótima desse problema, dentre os quais destacamos as técnicas conhecidas como estimadores S regularizados, que buscam minimizar uma função de perda dependente dos dados combinada com um termo de regularização. A introdução desse termo suscita um controle da capacidade de aproximação que pode trazer benefícios importantes tanto em termos de desempenho de generalização quanto de estabilidade numérica.

2.1. Modelos Regularizados

2.1.1. LASSO

A regularização LASSO tem por base uma penalização construída sobre a norma L_1 do vetor de parâmetros do modelo. O uso dessa norma não permite, via de regra, que a solução ótima seja obtida em forma fechada, mas favorece a obtenção de soluções esparsas, o que pode engendrar um processo subjacente de seleção de características [Hastie et al., 2009]. Matematicamente, o problema a ser resolvido se torna:

$$W_{lasso} = \arg\min_{W}\{||Y - WX||_2^2 + \lambda||W||_1\}, \tag{1}$$

onde W é o vetor de parâmetros, X representa a matriz de entrada composto pelos atrasos da série temporal e o parâmetro λ regula a esparsidade do modelo, sendo comumente determinado por meio de técnicas de validação cruzada. À medida que o valor de λ aumenta, mais variáveis são eliminadas

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

do modelo. Além da instância clássica do LASSO, existem variantes, como o LARS [Tibshirani et al., 2011], que realizam o processo de ajuste dos parâmetros de maneira mais eficiente.

2.1.2. RIDGE

A regularização *ridge* penaliza a soma dos quadrados dos parâmetros em vez de seus módulos, ou seja, tem por base a norma L_2. Nesse caso, é possível manter o problema no mesmo contexto quadrático da função custo, tornando mais simples o processo de diferenciação / determinação do gradiente e, no caso linear, permitindo a obtenção de uma solução em forma fechada [Hastie et al., 2009]. Matematicamente, tem-se:

$$W_{ridge} = \arg\min_{W}\{||Y - WX||_2^2 + \lambda||W||_2^2\} \tag{2}$$

Por sua vez, a Regularização de Kernel Ridge é uma extensão do modelo Ridge que incorpora o "truque do kernel" [Zhang et al., 2013]:

$$W_{kernel} = \arg\min_{W}\{||Y - W_{t,j}\boldsymbol{K}(X_t, X_j)||_2^2 + \lambda||W_{t,j}\boldsymbol{K}(X_t, X_j)||_2^2\} \tag{3}$$

onde $\hat{f}(X) = \sum_{t=1}^{T} W_{t,j}\boldsymbol{K}(X_t, X_j)$ é a função de previsão, $\boldsymbol{K}(X_t, X_j)$ é a função de kernel que quantifica a similaridade entre X_t (amostra de treinamento) e X_j (nova amostra a ser prevista), e $||\cdot||_2$ denota a norma L_2.

2.1.3. Elastic Net

A regularização Elastic Net é uma técnica que combina os termos de penalização L_1 e L_2. Dessa forma, busca-se obter benefícios de ambas as abordagens, mas enfrenta-se o desafio de determinar dois hiperparâmetros para alcançar soluções ótimas [Hastie et al., 2009].

$$W_{net} = \min_{W}\{||Y - WX||_2^2 + \lambda[\alpha||W||_1 + (1-\alpha)||W||_2^2]\} \tag{4}$$

Note que o uso dos métodos aqui descritos depende da seleção dos hiperparâmetros λ e α, a qual é realizada, via de regra, por meio de validação cruzada.

2.2. Métodos Bagging e Boosting

O método *Bagging* é uma técnica baseada em *bootstrap* que gera múltiplas amostras do conjunto de treinamento, treinando cada modelo preditor com uma redistribuição aleatória dessas amostras [Ribeiro e dos Santos Coelho, 2020], [Zhou, 2012]. Cada modelo preditor é treinado com um subconjunto aleatório, com reposição, das amostras originais, onde o tamanho desse subconjunto é igual ao tamanho do conjunto de treinamento original. Isso significa que algumas amostras originais podem ser selecionadas várias vezes, enquanto outras podem ser deixadas de fora. Cada modelo preditor no conjunto é gerado a partir de uma amostragem aleatória diferente do conjunto de treinamento. Ao realizar essa nova amostragem com substituição, algumas amostras são representadas várias vezes, enquanto outras são excluídas causando erro individual alto, o que, em conjunto, tende a resultar em erros no conjunto de testes inferiores aos de um único preditor.

Por outro lado, o método *Boosting* cria uma série de modelos preditores, onde o conjunto de treinamento para cada membro da série é selecionado com base no desempenho do preditor anterior. Nesse caso, as amostras que foram previstas incorretamente pelos preditores anteriores têm uma probabilidade maior de serem escolhidas para treinar o próximo preditor do que as amostras que foram previstas corretamente. Isso permite que o Boosting gere novos preditores capazes de lidar melhor com amostras nas quais o desempenho atual é insatisfatório. Enquanto, no Bagging, os modelos preditores são treinados simultaneamente, no Boosting eles são treinados sequencialmente.

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

Isso implica que uma série de modelos é construída, e, a cada iteração subsequente, os pesos das amostras mal previstas no modelo anterior são aumentados. Essa realocação de pesos auxilia o algoritmo na identificação dos parâmetros que precisa ajustar para melhorar seu desempenho.

2.3. Redes Neurais Recorrentes

As redes recorrentes constituem uma categoria específica de arquitetura de rede neural que, através da presença de laços de realimentação, pode ser capaz de prover meios para uma mais efetiva identificação de padrões em sequências temporais de dados. Um exemplo de rede neural recorrente que vem recebendo expressiva atenção nos últimos anos advém do paradigma de *Long Short-Term Memory* (LSTM) [Greff et al., 2017]. As redes LSTM são capazes de armazenar informações em seu estado interno, lidando com representações de memórias de curto e longo prazo. Essa capacidade é viabilizada por três componentes fundamentais que regulam o funcionamento das células LSTM: o componente de esquecimento, o componente de entrada e o componente de saída. Esses componentes controlam o fluxo de informações essenciais para a previsão na rede.

O componente de esquecimento (*forget gate*) tem a função de decidir quais elementos devem ser retidos no vetor de estado de longo prazo $\mathbf{c}_{t-1}$ no instante t, baseando-se na entrada X_t e na saída $\mathbf{h}_{t-1}$. O componente de entrada (*input gate*) é responsável por determinar a magnitude das alterações a serem feitas nos elementos do vetor de estado. Finalmente, o componente de saída (*output gate*) decide quais partes do novo vetor de estado de longo prazo $\mathbf{c}_t$ serão utilizadas como saída da célula LSTM. Assim, o LSTM é alimentado pela entrada X_t e pela sua própria saída $\mathbf{h}_{t-1}$.

O modelo LSTM é treinado utilizando o algoritmo de retro-propagação do erro através do tempo (BPTT, do inglês *backpropagation-through-time*) [Haykin, 2009], com o objetivo de minimizar a função de custo durante o processo de aprendizado. O BPTT opera de modo que o modelo seja treinado por meio do cálculo dos erros da camada de saída até a camada de entrada. Esses cálculos possibilitam o ajuste adequado dos parâmetros do modelo.

Uma versão simplificada do LSTM é o modelo Gated Recurrent Unit (GRU) [Greff et al., 2017], que opera de maneira semelhante. No entanto, o GRU se distingue de outras redes por expor todo o estado de uma célula a outras unidades da rede, ao passo que uma célula LSTM controla a exposição do conteúdo de sua memória. Essas abordagens expandem todas as possibilidades de modelagem de dados com dependência temporal.

2.4. Redes Neurais Convolucionais Dilatadas

Uma rede neural convolucional dilatada (DCNN) é uma variante de uma CNN convencional que incorpora camadas de convolução dilatada, também conhecida como convolução com espaçamento [Borovykh et al., 2018]. A DCNN possui a capacidade de eficientemente capturar padrões em dados sequenciais com diferentes escalas de tempo, incluindo padrões não lineares, quebras estruturais e longa memória. Os elementos centrais do modelo DCNN incluem o número de camadas convolucionais, a taxa de dilatação d, o tamanho dos filtros k, a direção do passado $s-d\cdot j$ (Eq. 5). Ao utilizar uma camada dilatada em uma CNN para tarefas de previsão de preços da soja, as informações temporais são capturadas em diversas escalas, tornando o modelo robusta na identificação de padrões nos dados. Para uma entrada de sequência $1D$, $X \in \mathbb{R}^n$ e um filtro $f : \{0, \cdots k-1\} \rightarrow \mathbb{R}$, a operação de convolução dilatada D no elemento s da sequência é definida como:

$$D(s) = (X *_d f)(s) = \sum_{j=0}^{k-1} f(j) \cdot X_{s-d\cdot j}, \tag{5}$$

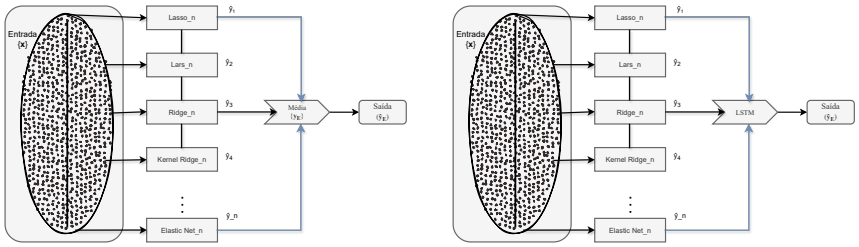
com $d = 1$, uma convolução dilatada é equivalente a uma convolução regular. O aumento da taxa de dilatação d permite que uma saída em níveis superiores represente uma variedade mais ampla de entradas. Neste estudo, foi adotada uma arquitetura DCNN composta por três camadas ocultas, conforme apresentada na Tabela 3.

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

3. Métodos Propostos

3.1. Comitês de Máquinas Heterogêneas

Os comitês de máquinas propostos para prever o preço da soja são compostos pela agregação das saídas de n modelos de regressão regularizados. A Fig. 1 mostra as estruturas adotadas para esses comitês. Na Figura 1(a), cada um dos n modelos ajusta os dados individualmente, e a previsão final é obtida pela média aritmética de suas saídas. Na Figura 1(b), os n modelos também aprendem de forma individual, e a previsão final é gerada por um modelo LSTM, que é treinado com as saídas de cada um dos n modelos. O termo "heterogêneo"é usado devido à diversidade de modelos de regressão, como LASSO, LARS, Ridge, Kernel Ridge, Elastic Net e redes neurais LSTM, que compõem o comitê. A técnica de *grid search* é empregada para realizar uma busca exaustiva de todas as combinações possíveis de hiperparâmetros. O parâmetro de regularização λ é selecionado por meio da técnica de validação cruzada (CV).



(a) (b)

Figura 1: Estruturas de comitês: (a) Combinador "Média Aritmética"e (b) Combinador "Modelo LSTM".

Para alcançar esse objetivo, adotamos o método construtivo. Começamos com um conjunto vazio e gradualmente adicionamos os n modelos. Após cada adição, avaliamos se há uma melhoria no erro quadrático médio (MSE); caso contrário, interrompemos a inclusão de modelos no comitê. O objetivo é maximizar o desempenho de generalização selecionando os n modelos candidatos que compõem o comitê. Além disso, é importante que os perfis de erros dos n modelos candidatos sejam descorrelacionados. Embora os modelos candidatos possam não generalizar satisfatoriamente individualmente, a agregação deles tende a produzir uma generalização melhor. Por esse motivo, além de utilizar a média aritmética, também propomos um modelo LSTM como agregador robusto para potencializar ainda mais o desempenho. Observamos que as entradas do modelo LSTM são as saídas dos n modelos candidatos, conforme ilustrado na Figura 1(b).

3.2. Transformer

O Transformer [Vaswani et al., 2017], é um modelo de atenção que se destaca pela capacidade de paralelização. Ele é constituído por dois componentes principais: um bloco codificador e um decodificador, que são interconectados. O bloco codificador é composto por m codificadores, enquanto o bloco decodificador é formado por m decodificadores. Cada codificador tem a mesma estrutura, embora utilize pesos diferentes, e é dividido em duas subcamadas: uma de autoatenção e outra de rede *feedforward*. Da mesma forma, os decodificadores incluem essas duas subcamadas, mas com a adição de uma camada de atenção entre elas, que permite ao decodificador se concentrar nas partes mais relevantes da sequência de entrada. O mecanismo de atenção é calculado através do produto escalar entre os vetores de Query-Key-Value (Q,K,V) [Zeng et al., 2023]:

$$Att(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\boldsymbol{T}}}{\sqrt{d_k}}\right) \cdot \boldsymbol{V} \tag{6}$$

$$MHAtt(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(h_1, h_2, \cdots, h_H) \cdot \boldsymbol{W^o} \tag{7}$$

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

onde $h_j = Att(\boldsymbol{QW_j^Q}, \boldsymbol{KW_j^K}, \boldsymbol{VW_j^V})$, com $\boldsymbol{Q} \in \mathbb{R}^{N \cdot d_k}$, $\boldsymbol{K} \in \mathbb{R}^{M \cdot d_k}$, $\boldsymbol{V} \in \mathbb{R}^{M \cdot d_v}$, $MHAtt$ denota a Atenção Multi-Cabeça, d_k e d_v denotam as dimensões de consultas e valores e h_j denotam as cabeças. A rede feedforward (FF) introduz a não-linearidade e uma camada densa.

$$FF(\boldsymbol{H'}) = swish(\boldsymbol{H'W}^1 + \boldsymbol{\beta}^1)\boldsymbol{W}^2 + \boldsymbol{\beta}^2 \tag{8}$$

$$\boldsymbol{H'} = LayerN(selfAtt(\boldsymbol{X}) + \boldsymbol{X}) \tag{9}$$

$$\boldsymbol{H} = LayerN(FF(\boldsymbol{H'}) + \boldsymbol{H'}) \tag{10}$$

onde $selfAtt(\cdot)$ denota o módulo auto-atenção e $LayerN(\cdot)$ a operação de normalização da camada, $swish$ é a função de ativação não linear utilizada, que é a novidade deste proposto e $\boldsymbol{H'}$ saídas anteriores. O Transformer busca capturar as dependências temporais dos dados utilizando a função de autocorrelação ao longo do tempo.

4. Experimentos

Este estudo visa avaliar a eficácia dos métodos de comitês de máquinas heterogêneas, bem como da rede Transformer, na previsão do preço da soja em dólares americanos (USD) no mercado brasileiro para um horizonte temporal de 30 dias à frente. As séries temporais foram obtidas do www.cepea.esalq.usp.br/br/indicador/soja.aspx, cobrindo o período de 13 de março de 2006 a 30 de março de 2024, totalizando 4.500 observações (Figura 2).

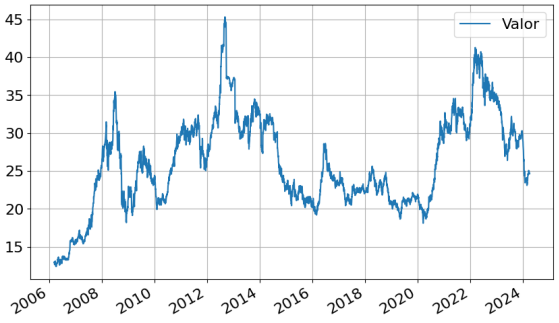


Figura 2: Evolução do preço da soja (USD).

Na Figura 2, podemos observar os picos na evolução do preço da soja ao longo do período de março de 2006 a março de 2024. O início desse intervalo mostra aumentos sucessivos no preço, chegando a um patamar em torno de 25.00 USD por sacas de 60 kg em 2006-2007. Em 2008, a crise financeira mundial impulsionou o preço da soja, seguido por outro aumento devido à crise política no Brasil em setembro de 2012, culminando com o maior pico no período no início de 2020. O período encerra com uma tendência de aumentos sucessivos no preço da soja, influenciada pela crise sanitária da COVID-19. A oscilação do preço da soja no Brasil entre março de 2006 e março de 2024 atingiu o valor máximo de 45.32 USD em 06 de setembro de 2012 e o valor mínimo de 12.4 USD em 10 de abril de 2006. A Tabela 1 fornece informações sobre as principais medidas exploratórias, como média, mediana, desvio padrão, assimetria e curtose da série temporal.

Tabela 1: Estatística da série.

Estatística	Obs.	Média	Des. Padrão	Máximo	Mínimo	Assimetria	Kurtosis
Valor	4.500	26.045	6.107	45.32	12.4	0.227	-0.21

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

A Tabela 1 revela que a série de preços da soja é assimétrica, enquanto a medida de curtose indica que a série de preços é platicúrtica. Ambas as medidas são empregadas para examinar se as amostras de dados são provenientes de uma distribuição normal. Essas duas medidas indicam que a série segue uma distribuição não normal.

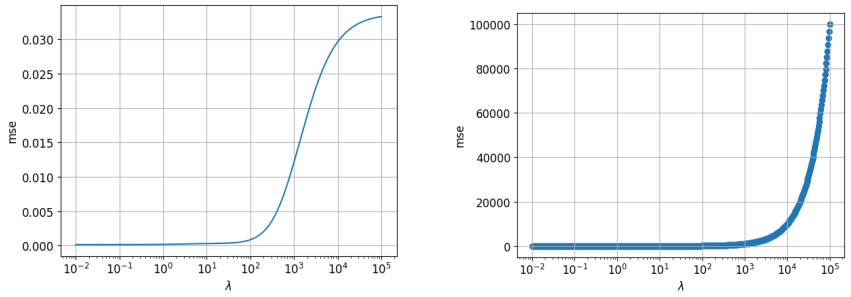
A validação cruzada foi utilizada para ajustar os métodos propostos neste estudo. Essa técnica envolve a aleatória divisão do conjunto de dados de treinamento em ”k”grupos. Durante k iterações, cada uma considera $k-1$ grupos para o ajuste dos parâmetros e um grupo para avaliar o desempenho do modelo. A Tabela 2 descreve os períodos utilizados na série de preços da soja.

Tabela 2: Divisão de Dados

Dado	Período	Amostra	Técnica
Conjunto de treinamento	13/03/2006 a 10/06/2022	4050.00	k-fold CV ($k \in [4, 10]$)
Conjunto de teste	13/06/2022 a 30/03/2024	445.00	h = 30 dias à frente

As simulações foram conduzidas utilizando o conjunto de dados de treinamento, onde diferentes defasagens foram consideradas para os dados de entrada. O conjunto de teste foi exclusivamente reservado para prever passos à frente. Os parâmetros de controle foram cuidadosamente ajustados para maximizar o desempenho, medido pelos erros quadráticos médios. Como resultado dessa análise, identificou-se a necessidade de utilizar cinco defasagens na previsão de múltiplos passos à frente, seguindo a estratégia direta. Além disso, os dados foram normalizados conforme o procedimento a seguir: $\mathbf{x}_T = \frac{x_t - min(x_t)}{max(x_t) - min(x_t)}$. O parâmetro ótimo λ é determinado com base nos erros quadráticos médios, como exemplificado na Figura 3. Foram avaliados diferentes valores de λ para cada modelo dentro do comitê.

Os comitês (Média e LSTM) compartilham os mesmos hiperparâmetros entre os 15 modelos que os compõem, diferindo apenas no combinador. O parâmetro ótimo λ é selecionado com base nos erros quadráticos médios, conforme ilustrado na Figura 3. O modelo Transformer proposto é relativamente simples devido ao tamanho das amostras disponíveis e contém 16.201 parâmetros treináveis. Enquanto isso, o modelo DCNN utiliza o otimizador ADAM, a função de ativação Swish para o ajuste dos pesos sinápticos e contém 21.645 parâmetros treináveis.



(a) (b)

Figura 3: Seleção de parâmetro λ: (a) Modelo Ridge e (b) Modelo Lasso.

Na Tabela 3, são apresentadas todas as escolhas de hiperparâmetros e seus respectivos valores adotados para os métodos propostos. Foi utilizada a técnica ”GridSearchCV”para testar todas as combinações e selecionar o modelo com melhor desempenho, definido como o mínimo erro quadrático médio. As combinações determinadas pela GridSearchCV são apresentadas nesta

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

tabela.

Tabela 3: Configurações e Hiperparâmetros adotados

Modelo	**Parâmetro/Hiperparâmetro**	**Valor/Técnica**
Comitês:	Modelos (Lasso, Lars, Ridge, Kernel Ridge, Elastic Net)	15
• *Média* &	Validação cruzada	$k \in [4, 10]$
	Máxima iteração	$i \in [500, 2000]$
	L_1 e L_2 penalização	$\lambda \in [10^{-2}, 10^{3}]$
• *LSTM*	Camada oculta	1
	Neurônios ocultos	50
	Taxa de aprendizagem	0.001
	Número de épocas	200
	função de ativação	ReLU
	Otimizador	ADAM
Transformer	Bloco Transformer	4
	Número de cabeças	4
	Neurônios ocultos	256
	Taxa de aprendizagem	10^{-4}
	Número de épocas	200
	Tamanho de lote	32
	Taxa de dropout	15%
	Otimizador	ADAM
	Função de ativação	Swish
	Parâmetros treináveis	16.201
DCNN	Camadas ocultas	3
	Neurônios ocultos	[100, 64, 10]
	Kernel size	[3, 3, 3]
	Taxa de dilatação	[1, 2, 4]
	Otimizador	ADAM
	Função de ativação	Swish
	Parâmetros treináveis	21.645
Bagging & Boosting	Base estimador	[XGBoost, Não]
	n_estimador	[20, 100]
	Bootstrap	[Não, Sim]
	Validação cruzada	k = 5
	Taxa de aprendizagem	0.05
	Número de combinação	[144, 192]

4.1. Abordagem Direta

A abordagem direta implica prever cada horizonte de tempo H de maneira independente. Em outras palavras, são ajustados H modelos separados (um para cada período de tempo) da série $\mathbf{y}_N$. Cada modelo é responsável por prever um único período, e as previsões são obtidas diretamente a partir do modelo estimado. Nesta abordagem, os erros para diferentes horizontes de previsão são independentes condicionalmente. Matematicamente, a abordagem direta pode ser definida como:

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

$$\hat{\mathbf{Y}}_{\mathbf{N+H}} = \begin{cases} \hat{y}_{N+1} = \Psi_1(\Theta, y_{N-\tau+1}), & se\ h = 1 \\ \hat{y}_{N+2} = \Psi_2(\Theta, y_{N-\tau+2}), & se\ h = 2 \\ \vdots & \\ \hat{y}_{N+H} = \Psi_H(\Theta, y_{N+H-\tau}), & se\ h = H \end{cases} \tag{11}$$

4.2. Métricas de Avaliação de Desempenho

As métricas a seguir foram empregadas para avaliar o desempenho dos métodos propostos: erro percentual médio absoluto (MAPE), erro quadrático médio (MSE), raiz do erro quadrático médio (RMSE) e a Eficiência Nash-Sutcliffe (NSE), cuja definição está apresentada abaixo:

$$MAPE\ (\%) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100 \tag{12}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{13}$$

$$RMSE = \left|\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right|^{\frac{1}{2}} \tag{14}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{15}$$

onde NSE $\in [-\infty, 1]$.

5. Resultados

Nesta seção, é realizada uma comparação do desempenho na previsão de diferentes passos à frente entre diversos modelos, incluindo os comitês propostos Média e LSTM como agregadores, Transformer, DCNN, Bagging e Boosting. Foi determinado o uso de cinco defasagens, e essa mesma ordem de defasagem foi aplicada a todos os modelos analisados. As previsões são conduzidas no conjunto de dados de teste utilizando a abordagem direta. Os resultados das métricas de avaliação de desempenho de todos os modelos utilizados neste experimento são apresentados na Tabela 4.

Os resultados apresentados na Tabela 4 indicam que o método de comitê com o modelo LSTM como agregador obteve o melhor desempenho em termos de NSE e MAPE para a previsão de um passo à frente. Para a previsão de cinco passos à frente, o modelo DCNN mostrou-se superior em termos de NSE e MAPE, com menor erro de previsão e maior qualidade. Entretanto, para previsões de 15 a 30 passos à frente, tanto o combinador Média quanto o LSTM apresentaram incertezas devido à volatilidade do preço da soja, resultando em um desempenho insatisfatório para os dois métodos propostos. Observamos que o método Boosting demonstrou um desempenho superior em termos de MAPE na previsão de 15 a 25 passos à frente. Em termos de qualidade, a rede Transformer demonstrou um melhor desempenho na previsão de 10 a 30 passos à frente. Observa-se que o NSE da rede Transformer apresenta um decaimento mais lento ao longo dos horizontes temporais em comparação com os demais modelos. É importante destacar que tanto a rede Transformer quanto o comitê-LSTM pertencem a um subconjunto de modelos considerados estatisticamente robustos. Esses resultados sugerem que esses modelos são capazes de lidar com a não estacionariedade e a longa dependência presentes na série de preços da soja, tornando-os escolhas robustas para essas características dos dados.

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

Tabela 4: Desempenho de Previsão

		Horizonte temporal						
Modelo	Métrica	h = 1	h = 5	h = 10	h = 15	h = 20	h = 25	h = 30
Comitê – Média	MAPE	0.0185	0.0443	0.0573	0.0685	0.0768	0.0894	0.0993
	MSE	0.0003	0.0010	0.0018	0.0026	0.0034	0.0048	0.0057
	RMSE	0.0165	0.0321	0.0424	0.0508	0.0579	0.0689	0.0758
	NSE	0.9753	0.9412	0.9045	0.8674	0.8120	0.7620	0.6948
Comitê – LSTM	MAPE	**0.0163**	0.0393	0.0542	0.0667	0.0792	0.0932	0.1044
	MSE	0.0002	0.0009	0.0017	0.0026	0.0037	0.0052	0.0064
	RMSE	0.0143	0.0297	0.0409	0.0508	0.0606	0.0719	0.0801
	NSE	**0.9898**	0.9528	0.9141	0.8706	0.8285	0.7816	0.7056
Transformer	MAPE	0.0191	0.0258	0.0402	0.0539	0.0639	0.0767	0.0935
	MSE	0.0004	0.0005	0.0009	0.0017	0.0024	0.0033	0.0049
	RMSE	0.0257	0.0264	0.0299	0.0409	0.0491	0.0576	0.0703
	NSE	0.9776	0.9709	**0.9548**	**0.9407**	**0.9316**	**0.9116**	**0.8869**
DCNN	MAPE	**0.0178**	**0.0225**	**0.0336**	0.0548	0.0612	0.0687	0.0797
	MSE	0.0002	0.0003	0.0006	0.0017	0.0021	0.0027	0.0036
	RMSE	0.0134	0.0171	0.0255	0.0413	0.0462	0.0520	0.0603
	NSE	0.9869	**0.9774**	0.9502	0.9258	0.8864	0.8648	0.8401
Bagging	MAPE	0.0210	0.0291	0.0468	0.0615	0.0737	0.0822	0.0999
	MSE	0.0003	0.0004	0.0011	0.0019	0.0027	0.0033	0.0048
	RMSE	0.0153	0.0212	0.0334	0.0434	0.0519	0.0577	0.0694
	NSE	0.9813	0.9654	0.9240	0.8853	0.8236	0.7457	0.6740
Boosting	MAPE	0.0213	0.0241	0.0373	**0.0462**	**0.0530**	**0.0617**	0.0747
	MSE	0.0003	0.0003	0.0008	0.0013	0.0017	0.0024	0.0036
	RMSE	0.0167	0.0186	0.0289	0.0359	0.0413	0.0489	0.0604
	NSE	0.9878	0.9734	0.9359	0.9009	0.8691	0.8179	0.7250

Nota: (1) o valor em negrito indica o melhor desempenho de previsão.

6. Conclusões

Neste estudo, foram avaliados métodos de comitês de máquina e redes neurais de última geração para prever os preços da soja com múltiplos passos à frente, utilizando a abordagem direta de previsão. O conjunto de dados abrangeu o período de 13/03/2006 a 30/03/2024 e revelou três características principais: não estacionariedade, quebras estruturais e longas memórias. Os resultados mostraram que o comitê com o modelo LSTM como agregador apresentou menores NSE e MAPE para previsões de curto prazo, enquanto o modelo DCNN teve melhor desempenho na previsão de cinco passos à frente. A rede Transformer demonstrou uma qualidade superior ao longo do horizonte temporal, com um NSE que diminuiu lentamente ao longo do tempo. No geral, os comitês de máquina propostos neste estudo obtiveram resultados satisfatórios, embora com um desempenho ligeiramente inferior em comparação com outros modelos. Concluímos que tanto a rede Transformer quanto o comitê-LSTM são considerados modelos estatisticamente robustos, capazes de lidar com as características específicas dos dados, e podem ser uma escolha sólida para previsões de séries temporais com esses padrões.

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br

Agradecimento
O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) e CNPq no. 308425/2023-5 e no âmbito do Centro de Pesquisa Aplicada em Inteligência Artificial BIOS - Brazilian Institute of Data Science, apoiado pelo processo no. 2020/09838-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Referências

Borovykh, A., Bohte, S., e Oosterlee, C. W. (2018). Dilated convolutional neural networks for time series forecasting. *Journal of Computational Finance, Forthcoming*.

CONAB (2024). Análises do Mercado Agropecuário e Extrativista - Boletim Logístico. Acessado: 2024-05-16.

Divina, F., Gilson, A., Goméz-Vela, F., Torres, M. G., e Torres, J. F. (2018). Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting. *Energies*, 11(4):1–31.

FAO (2024). Food and Agriculture Organization of the United Nations (FAO). FAO database. Acessado: 2024-03-29.

Greff, K., Srivastava, R., Koutnik, J., Steunebrink, B., e Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10): 2222–2232. ISSN 2162-237X, 2162-2388.

Hastie, T., Tibshirani, R., e Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer. ISBN 9780387952840.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Pearson International Ed. Pearson.

Liu, Y., Duan, Q., Wang, D., Zhang, Z., e Liu, C. (2020). Prediction for hog prices based on similar sub-series search and support vector regression. *Computers and Electronics in Agriculture*, 157: 581–588. ISSN 0168-1699.

Ribeiro, M. H. D. M. e dos Santos Coelho, L. (2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Applied soft computing*, 86:105837.

Tibshirani, R., Taylor, J., Candès, E., e Hastie, T. (2011). *The Solution Path of the Generalized Lasso*. Stanford University.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., e Polosukhin, I. (2017). Attention is all you need. In *Neural Information Processing Systems*.

Zeng, A., Chen, M., Zhang, L., e Xu, Q. (2023). Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on AI*, volume 37, p. 11121–11128.

Zhang, Y., Duchi, J. C., e Wainwright, M. J. (2013). Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal Mach. Learn. Res.*, p. 3299–3340.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., e Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, p. 11106–11115.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC.

DOI: 10.59254/sbpo-2024-193665
https://proceedings.science/p/193665?lang=pt-br