FACULTAD DE INGENIERÍA
REVISTA DE LA FACULTAD DE INGENIERIA
www.revistaingenieria.uda.cl

Modelos de previsão de preços no mercado de combustíveis

Ricardo C. A. Araujo[1,2] and Paulo H. Ferreira[3]

[1]Secretaria da Fazenda e Planejamento do Estado de São Paulo, Avenida Rangel Pestana 300, Sé, CEP: 01017-911, São Paulo, SP, Brasil; *rcaaraujo@outlook.com.*
[2]MBA em Ciências de Dados, CeMEAI, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Avenida Trabalhador São-Carlense 400, Centro, CEP: 13.566-590, São Carlos, SP, Brasil.
[3]Departamento de Estatística, IME, Universidade Federal da Bahia, Avenida Milton Santos s/n, Campus de Ondina, CEP: 40.170-110, Salvador, BA, Brasil; *paulohenri@ufba.br.*

RESUMO

Os combustíveis fósseis permanecem como uma das principais fontes de energia. É de suma importância para o setor público, na determinação de sua política fiscal e tributária, uma estimativa do comportamento do preço de determinado combustível. Com base em conhecimentos empíricos adquiridos no acompanhamento do mercado, acredita-se em uma significativa influência da gasolina sobre os preços praticados para o etanol combustível. O nível de produção e a quantidade de estoque do etanol são definidos, basicamente, enquanto a diferença de preços desses dois combustíveis seja capaz de oferecer uma atraente margem de lucro aos produtores de etanol. Este trabalho propõe contribuir para o processo de tomada de decisão da Secretaria da Fazenda e Planejamento do Estado de São Paulo com o desenvolvimento de ferramentas para predição do comportamento do preço de determinado produto com base em seus produtos concorrentes ou similares. Mais especificamente, o objetivo do presente estudo é predizer os preços de revenda do etanol com base na série histórica dos preços médios semanais da gasolina.

Palavras-chaves: etanol; gasolina; hiperparâmetro; preço; previsão.

ABSTRACT

Fossil fuels remain one of the primary sources of energy. It is of utmost importance for the public sector, in determining its fiscal and tax policy, to have an estimate of the price behavior of a given fuel. Based on empirical knowledge gained from monitoring the market, it is believed that gasoline significantly influences the prices applied to ethanol fuel. The production level and stock quantity of ethanol are primarily determined by whether the price difference between these two fuels offers an attractive profit margin to ethanol producers. This study aims to contribute to the decision-making process of the São Paulo State Department of Finance and Planning by developing tools to predict the price behavior of a given product based on its competing or similar products. More specifically, the objective of this study is to predict ethanol resale prices based on the historical series of weekly average gasoline prices.

Palavras-chaves: ethanol; gasoline; hyperparameter; price; prediction.

1 Introdução

Apesar do crescente nível de investimentos no mercado de veículos elétricos, os combustíveis fósseis derivados de petróleo ainda desempenham um importante papel na economia global, permanecendo como uma das principais fontes de energia para o setor de transporte de mercadorias e pessoas.

Assim como no setor privado, é de suma importância para o setor público, na concessão de benefícios fiscais e na determinação de sua política fiscal e tributária, uma estimativa do comportamento do preço de determinado produto.

Segundo Zhang & Zhao (2022), como um dos mais importantes produtos derivados do petróleo, a demanda de gasolina é constantemente afetada por diversos fatores. Desde o crescimento do Produto Interno Bruto (PIB), do impacto inflacionário, da redução na emissão de carbono e do desenvolvimento da indústria automotiva e de transportes, variados e influentes fatores afetam a demanda desse combustível. Com frequentes flutuações, a previsão de sua demanda torna-se complexa e extremamente difícil. Na literatura existente, uma grande variedade de modelos econométricos e de aprendizado de máquina têm sido utilizados para a criação de modelos para previsão do consumo e da demanda no mercado de energia. Por exemplo, Azadeh *et al.* (2008) primeiramente propuseram uma rede neural para prever o consumo a longo prazo de eletricidade em grandes consumidores industriais e demonstraram as vantagens da abordagem da rede neural por meio de uma análise de variância (ANOVA). Yan *et al.* (2019) combinaram redes neurais *Long Short-Term Memory* (LSTM) com técnicas de *Stationary Wavelet Transform* (SWT) para propor um modelo híbrido de *deep learning* para previsão do consumo doméstico de energia. Os resultados experimentais demonstraram maior eficiência no treinamento do método proposto, com ganhos de precisão nos valores previstos. Yu *et al.* (2021) adotaram um eficaz modelo de conjunto de *Rolling Decomposition-Ensemble* (RDE) para previsão do consumo da gasolina e obtiveram previsões com melhor desempenho que os modelos de *benchmark* listados no estudo. Similarmente, Yu & Ma (2021) propuseram um modelo de conjunto de RDE orientado ao tratamento dos dados para previsão do consumo da gasolina e obtiveram previsões com bons desempenhos.

Li *et al.* (2022) apontam na atual literatura uma importante limitação na aplicação no mercado de energia dos modelos de aprendizado de máquina quando, sob uma perspectiva da ciência da computação, parâmetros são enfatizados enquanto uma interpretação econômica/financeira pode ser ignorada. Por exemplo, dados que impactam a taxa de precisão podem ser tratados como ruído apesar de seu possível significado econômico. Por outro lado, outros estudos concentrados num enfoque econômico/financeiro falham numa completa exploração da capacidade de seus algoritmos. Allen (2019) promove a visão de que os modelos de aprendizado de máquina podem acelerar uma crise na ciência pois, sendo seus algoritmos desenvolvidos especificamente para identificar informações em um conjunto de dados, ao realizarem uma busca em uma enorme massa de dados, inevitavelmente irão identificar algum padrão.

O intuito mais geral deste trabalho é, portanto, desenvolver um modelo de aprendizado de máquina para previsão dos preços do etanol com base numa série histórica de preços da gasolina. Devido à sua grande importância socioeconômica e com base em conhecimentos empíricos adquiridos no acompanhamento do mercado de combustíveis, acredita-se que os preços do etanol praticados pelo mercado sofram uma significativa influência dos preços praticados para a gasolina. Como objetivo mais específico, pretende-se contribuir para o processo de tomada de decisão da Secretaria da Fazenda e Planejamento do Estado de São Paulo com o desenvolvimento de ferramentas para predição do comportamento do preço de determinado produto baseadas na adoção de produtos concorrentes ou similares como parâmetros externos (hiperparâmetros).

O restante deste documento está organizado da seguinte forma. Apresenta-se na Seção 2 uma breve bibliografia contendo os modelos comumente utilizados. Na Seção 3, detalha-se a metodologia empregada na previsão da demanda e avaliação dos erros. Em seguida, na Seção 4, mostram-se os dados utilizados e os resultados obtidos. Por fim, na Seção 5, apresentam-se as conclusões do trabalho, futuras melhorias e aplicações para a modelagem.

2 Revisão Bibliográfica

Segundo Ghoddusi *et al.* (2019), a performance superior dos modelos de aprendizado de máquina

no processamento, classificação e previsão de informações complexas e de larga escala os tornou populares em muitas áreas da indústria de energia, sendo largamente utilizados em aplicações relacionadas à análise econômica e financeira. Uma lista exemplificativa inclui estudos na área de exploração de óleo e gás (Anifowose *et al.*, 2017), no processamento de óleo e gás (Zendehboudi *et al.*, 2018), na previsão da capacidade de geração de energia solar (Voyant *et al.*, 2017), na otimização de geradores de energia (Zeng *et al.*, 2018), na previsão da capacidade de geração de energia eólica (Heinermann & Kramer, 2016), na otimização de geradores de energia eólica (Marugán *et al.*, 2018), na previsão de falhas (Gupta *et al.*, 2015), na previsão da demanda de carga (Jurado *et al.*, 2015) e na geração de energia hidráulica (Zaidi *et al.*, 2018).

Uma comparação entre as características dos modelos de aprendizado de máquina com os modelos econométricos tradicionais (por exemplo, *Autoregressive Integrated Moving Average* (ARIMA) e *Generalized Autoregressive Conditional Heteroskedasticity* (GARCH)) revela algumas das razões para o aumento de sua popularidade. Algoritmos de aprendizado de máquina podem manipular dados estruturados ou não em uma larga escala, permitindo rapidamente a tomada de decisões e previsões. Tal superioridade é possível devido à ausência de qualquer presunção prévia acerca do tipo de função equacional, da interação entre as variáveis e da distribuição estatística dos parâmetros. No aprendizado de máquina, variáveis de saída são previstas com base em outras variáveis.

Zhang & Zhao (2022) propõem um modelo de decomposição de séries temporais orientado pelas características de tendência e periodicidade da demanda de gasolina no mercado chinês. Resultados empíricos demonstram que modelos de decomposição temporal com tendência linear e sazonalidade é uma solução viável para previsão da demanda de gasolina em mercados com traços de tendência e periodicidade.

Santos (2021) propõe a aplicação de uma ferramenta que busca a otimização do lucro bruto pela metodologia da precificação dinâmica suportada pela inteligência artificial, modelos de aprendizado de máquina e cálculos logarítmicos. Com o objetivo central de analisar seu desempenho comparado com o gerenciamento tradicional de preços, conclui-se que a análise traz à tona desafios gerenciais no que tange à aplicabilidade da ferramenta e do método não só à revenda varejista de combustíveis, mas também àqueles negócios que buscam se destacar por uma abordagem profissional ao preço do seu produto e/ou serviço disponibilizado. A manipulação da precificação usando aprendizado de máquina não se mostrou superior ao modelo tradicional dentro do período analisado. Teoricamente, a modelagem de aprendizado de máquina do exemplo analisado poderia buscar parâmetros qualitativos, inseridos num aprendizado supervisionado que, através de algoritmos de classificação, alcançariam o melhor posicionamento de preços perante a concorrência.

Beyca *et al.* (2019) propõem o emprego de três modelos populares de aprendizado de máquina para previsão do consumo de gás natural na província de Istambul, Turquia. Adotados os modelos de regressão linear múltipla, redes neurais e modelo vetorial autorregressivo (VAR), os resultados indicam superioridade do modelo VAR sobre as técnicas de redes neurais, fornecendo resultados mais confiáveis e precisos para previsão de séries temporais do consumo de gás natural.

Almeida *et al.* (2016) propõem estimar um sistema de demanda para gasolina comum, etanol hidratado e óleo diesel via modelo *Linear Approximation Almost Ideal Demand System* (LA-AIDS) com dados de séries temporais trimestrais para o período de 2001 a 2015 no estado brasileiro de Pernambuco. As estimações foram feitas por meio do método *Seemingly Unrelated Regressions* (SUR). Os resultados encontrados são semelhantes aos observados na literatura, no sentido em que apontam para a inelasticidade-preço da demanda de gasolina e diesel, o que é esperado dada a essencialidade dos bens. Também foi possível verificar que a demanda por etanol é elástica por ter elasticidade-preço Marshalliana maior que um em valor absoluto, convergindo para os resultados já encontrados em outros estudos.

Melo (2012) analisa a relação entre o mercado de gasolina e o mercado de etanol. Através do modelo VAR, procurou-se entender o efeito da substituição entre os combustíveis, notando-se no curto prazo um maior efeito do preço da gasolina na demanda de etanol. No longo prazo, porém, os consumidores aumentam a demanda pelo biocom-

bustível. A escolha do consumidor pela gasolina é predominante no curto prazo. No entanto, um aumento repentino no preço do combustível fóssil leva a uma substituição de combustíveis ao longo do tempo.

Sobreiro *et al.* (2008) comparam o desempenho das redes neurais artificiais (RNAs) usando a arquitetura *Perceptron* multicamadas, com o do método ARIMA para previsão do preço do etanol combustível. Como resultado, observou-se que a aplicação das RNAs obteve uma aproximação mais satisfatória quando comparada à aplicação do método ARIMA, o que em conclusão evidencia a importância das RNAs na previsão dos preços do etanol combustível.

3 Metodologia

Utilizando uma base de dados real, o objetivo do presente estudo é estimar o mercado do etanol para tomada de decisão na concessão de benefícios fiscais e na determinação de políticas fiscais e tributárias, dada a série histórica dos preços médios semanais da gasolina desde maio de 2004 (últimos 18 anos) no estado de São Paulo, Brasil.

Primeiramente, com base na metodologia *Cross-Industry Standard Process for Data Science* (CRISP-DS), o desenvolvimento do projeto será realizado obedecendo às seguintes etapas, conforme ilustrado na Figura 1:

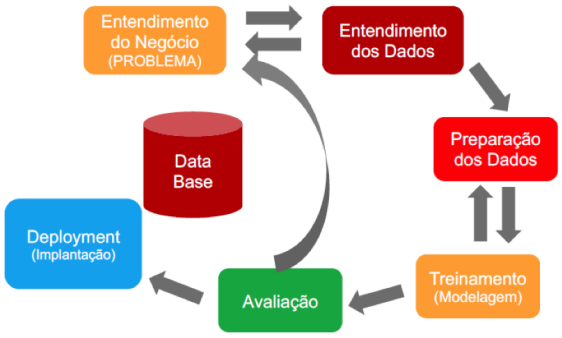


Figura 1: *Workflow* da metodologia CRISP-DS. Fonte: Material da aula de Gestão de Projeto de Ciências de Dados do Prof. Francisco Louzada Neto.

1. Entendimento do negócio (definição do problema);
2. Entendimento dos dados;
3. Preparação dos dados;
4. Treinamento (modelagem);
5. Avaliação;
6. *Deployment*.

3.1 Entendimento do Negócio (Definição do Problema)

3.1.1 Etanol: Descrição e Mercado

O etanol, comercialmente chamado de álcool etílico, é um composto orgânico da família dos álcoois. Classificado de acordo com a concentração de água, possui dois tipos básicos: o anidro e o hidratado. Pela lei brasileira, no etanol hidratado é permitida uma maior concentração máxima de água.

Além de sua utilização como biocombustível, o etanol é amplamente adotado como insumo industrial em distintos setores da economia, desde a fabricação de detergentes, produtos de limpeza, pinturas e solventes à fabricação de perfumes, desodorantes, cremes e produtos de higiene em geral. Destaca-se ainda no setor alimentício e farmacêutico com a fabricação de bebidas, vinagre, vacinas, antibióticos e antissépticos. Especialmente, com a pandemia de COVID-19 (sigla do inglês, *Coronavirus Disease 2019*), o uso de álcool 70% para a higienização das mãos e superfícies obteve relevância na ajuda para prevenção ao coronavírus.

Segundo Schutte & Barros (2010), o etanol representa hoje mais de 90% do fornecimento mundial de biocombustíveis líquidos. É produzido, fundamentalmente, a partir da cana-de-açúcar e do milho, embora se possam utilizar outros cultivos amiláceos (Fao, 2009). O crescente interesse no etanol está relacionado diretamente ao aumento da preocupação com a degradação do meio ambiente, à busca de fontes de energia renováveis, à procura de uma diversificação destas fontes por motivos geopolíticos e à geração de oportunidades de trabalho e renda no campo. A produção mundial de etanol quase quadruplicou entre 2000 e 2008 (Fao, 2009). O Brasil e os Estados Unidos da América são os principais produtores, seguidos por China, Índia e França. O comércio internacional representa pouco mais de 10% da produção,

sendo o Brasil responsável por quase dois terços das exportações.

No mercado de combustíveis, devido ao seu poder calorífico menor que o da gasolina, estima-se que, para manter-se competitivo, o preço de varejo do etanol deve ser inferior a 70% do preço do litro da gasolina. Por consequência, variações no preço internacional do barril de petróleo interferem na demanda por alternativas mais baratas de combustíveis. Normalmente, o aumento do preço da gasolina faz com que o produtor do etanol combustível também eleve seu preço.

O preço do etanol também está atrelado à cotação de outras *commodities* no mercado internacional. Valorizações do preço externo do açúcar, outro importante produto extraído da cana-de-açúcar, também podem impactar no nível de produção do etanol.

Impreterivelmente, as safras anuais de cana-de-açúcar são outro importante fator de influência no preço do etanol. Adversidades climáticas, como períodos de seca e estiagem, impactam diretamente a área de cultivo e o nível de produção de sua matéria-prima.

Introduzido na matriz energética brasileira no início do século XX, como uma tentativa da redução de dependência do petróleo importado, o etanol somente obteve participação no mercado de combustíveis com o lançamento do Programa Nacional do Álcool (Proálcool) em 1975. Devido à sua baixa competitividade, contudo, foi mantido basicamente por meio de subsídios governamentais. Com o fim do regime militar e a forte crise econômica então enfrentada, os subsídios foram reduzidos e o etanol perdeu sua participação no mercado. O Brasil voltou a ser um grande exportador de açúcar.

O etanol somente veio a figurar como competidor da gasolina com a introdução dos veículos *flex fuel*. O desenvolvimento de novas tecnologias, simultaneamente com a redução dos custos de produção e o aumento internacional do valor do petróleo, dariam ao etanol o papel definitivo como alternativa aos combustíveis fósseis.

3.1.2 Transformar a Série Histórica de Preços em um Problema de Predição

Com base em conhecimentos empíricos adquiridos no acompanhamento do mercado de combustíveis, acredita-se que haja significativa influência do mercado da gasolina sobre os preços praticados para o etanol combustível. Seu nível de produção e a quantidade de estoque a ser disponibilizada no mercado interno são definidos, basicamente, considerando-se que a diferença entre o preço da gasolina e o preço do etanol seja capaz de oferecer uma atraente margem de lucro aos produtores de etanol (empresas usineiras de açúcar e álcool).

O objetivo do presente estudo é predizer os preços de revenda do etanol com base na série histórica dos preços médios semanais da gasolina desde maio de 2004 (últimos 18 anos) no estado de São Paulo.

Na Figura 2 é exibida a representação gráfica com a evolução dos preços médios semanais da gasolina e do etanol, objeto do presente estudo.

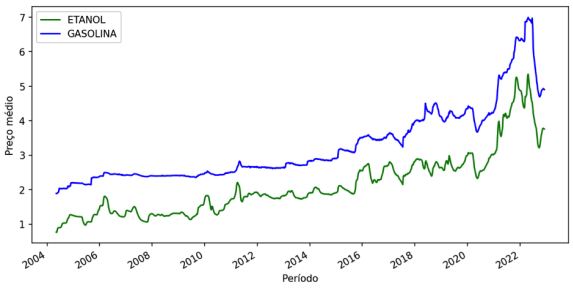


Figura 2: Série histórica dos preços médios semanais da gasolina e do etanol.

3.1.3 Perguntas a Serem Respondidas

Com a conclusão do presente estudo, pretende-se responder às seguintes perguntas:

- Existe relação confiável entre os preços da gasolina e do etanol combustível?
- Pode-se inferir que os produtores de etanol combustível adotam como referência o preço da gasolina?

3.2 Entendimento dos Dados

3.2.1 Coleta dos Dados

Para a realização do presente estudo, o conjunto de dados terá como base a série histórica de preços médios semanais fornecida pelo Sistema de Levantamento de Preços da Agência Nacional

do Petróleo (ANP)[1] entre maio/2004 e dezembro/2022.

3.2.2 Dicionário e Estrutura do Conjunto de Dados

Nesse conjunto de dados, existem 135.634 linhas e 18 colunas. A estrutura geral organiza as seguintes informações:

- Dia/mês/ano de **início e fim do período semanal** entre maio/2004 e dezembro/2022;
- **Região e estado** de referência;
- **Produto** (etanol hidratado, gasolina comum, gasolina aditivada, gás liquefeito de petróleo (GLP), gás natural veicular (GNV), óleo diesel e óleo diesel S10);
- **Número de postos pesquisados**;
- **Unidade de medida**;
- **Preço médio de revenda** (preço ao consumidor final), expresso em reais (R$), entre maio/2004 e dezembro/2022;
- **Cinco parâmetros estatísticos** do preço médio de revenda;
- **Preço médio de distribuição** (preço aos postos de combustível), expresso em R$, entre maio/2004 e dezembro/2022;
- **Quatro parâmetros estatísticos** do preço médio de distribuição.

3.2.3 Identificação dos Tipos das Variáveis

Nesta subsubseção são identificados os tipos de variáveis do conjunto de dados. Posteriormente, essas variáveis deverão ser restruturadas para uma melhor manipulação de suas informações.

- As variáveis associadas à região, estado e produto são **categóricas politômicas**;
- As variáveis dia/mês/ano são do tipo **data**;
- As variáveis associadas aos preços de revenda e de distribuição são do tipo **numérico contínuo**;
- As variáveis associadas ao número de postos são do tipo **numérico discreto**.

3.2.4 Análise Descritiva e Gráfica das Séries

Após consolidação e reestruturação de suas informações, o conjunto de dados apresenta a seguinte estrutura mostrada pelas Tabelas 1 e 2.

Ressalta-se que, com a eliminação das informações referentes aos produtos gasolina aditivada, GLP, GNV, óleo diesel e óleo diesel S10, e com a reorganização das informações referentes ao etanol hidratado e à gasolina comum em duas colunas distintas, o número total de linhas foi reduzido para 956 linhas.

Tabela 1: Conjunto de dados em análise.

Data_Inicial	Etanol	Gasolina
2004-05-09	0,768	1,891
2004-05-16	0,766	1,888
2004-05-23	0,823	1,894
2004-05-30	0,887	1,912
2004-06-06	0,894	1,919
⋮	⋮	⋮
2022-11-06	3,720	4,900
2022-11-13	3,770	4,910
2022-11-20	3,780	4,930
2022-11-27	3,770	4,920
2022-12-04	3,760	4,900

Tabela 2: Estatísticas do conjunto de dados.

	Etanol	Gasolina
n	956	956
Mínimo	0,766	1,888
1º Quartil	1,367	2,429
Média	2,124	3,240
Mediana	1,879	2,755
3º Quartil	2,623	3,904
Máximo	5,348	6,992
Desvio-padrão	0,922	1,094

Conforme observado na Tabela 2, as séries históricas de ambos os combustíveis apresentam uma distribuição com desvio-padrão pequeno e com valores próximos a 1.

A série histórica do etanol apresenta o seguinte comportamento de autocorrelação representado graficamente pelas Figuras 3 e 4. Pela análise quantitativa do grau de autocorrelação da série,

[1]Sistema de Levantamento de Preços da ANP. Disponível em: `https://www.gov.br/anp/pt-br/assuntos/precos-e-defesa-da-concorrencia/precos`. Acesso em: 17 de dezembro de 2022.

o coeficiente de autocorrelação calculado ($lag = 1$) é de 0,998.

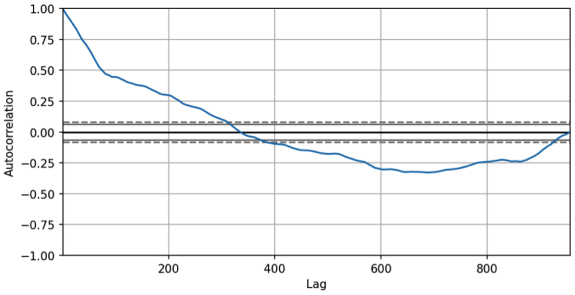


Figura 3: Gráfico de autocorrelação do etanol.

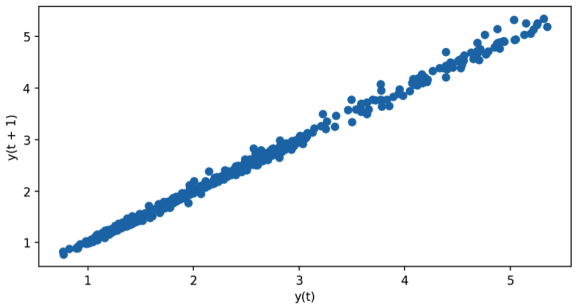


Figura 4: Gráfico de *lag plot* do etanol.

Ressalta-se que, na Figura 3, as linhas horizontais correspondem ao intervalo de confiança entre 95% e 99%. As linhas tracejadas correspondem ao intervalo de 99%. Sabendo-se que as observações localizadas fora desses intervalos de confiança são consideradas estatisticamente correlacionadas, é possível concluir que a curva de autocorrelação apresenta correlação por quase toda a sua extensão.

A série histórica da gasolina apresenta o seguinte comportamento de autocorrelação representado graficamente pelas Figuras 5 e 6. Pela análise quantitativa do grau de autocorrelação da série, o coeficiente de autocorrelação calculado ($lag = 1$) é de 0,999.

Assim como na Figura 3, pela Figura 5 é possível concluir que a curva de autocorrelação apresenta correlação por quase toda a sua extensão.

Conforme observado nas Figuras 3-6 em conjunto, as séries históricas apresentam um comportamento de autocorrelação bastante similar, observando-se uma autocorrelação positiva e decrescente até um *lag* de aproximadamente 300 semanas.

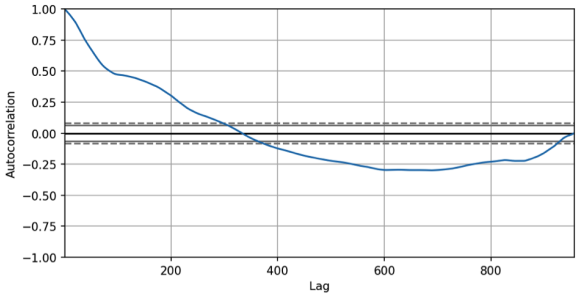


Figura 5: Gráfico de autocorrelação da gasolina.

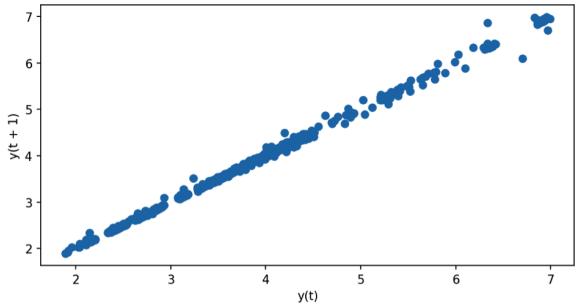


Figura 6: Gráfico de *lag plot* da gasolina.

Em particular, no gráfico de dispersão demonstrado na Figura 7, o agrupamento dos pontos em torno da linha de tendência reta demonstra o forte grau de associação linear positiva entre as séries. Ressalta-se que uma linha de tendência reta significa que o gráfico de dispersão é linear. Pela análise quantitativa do grau de associação linear entre o conjunto de dados, o coeficiente de correlação calculado (método de Pearson) é de 0,978.

Pelas análises até então realizadas, depreende-se que as séries históricas do etanol e da gasolina apresentam um comportamento estatístico muito similar, com uma forte correlação linear positiva entre si.

Por consequência, para os questionamentos levantados pelo presente estudo, estima-se que, de fato, há uma forte relação entre os preços da gasolina e do etanol. Provavelmente, os preços de mercado da gasolina possuem relevante influência sobre os produtores de etanol.

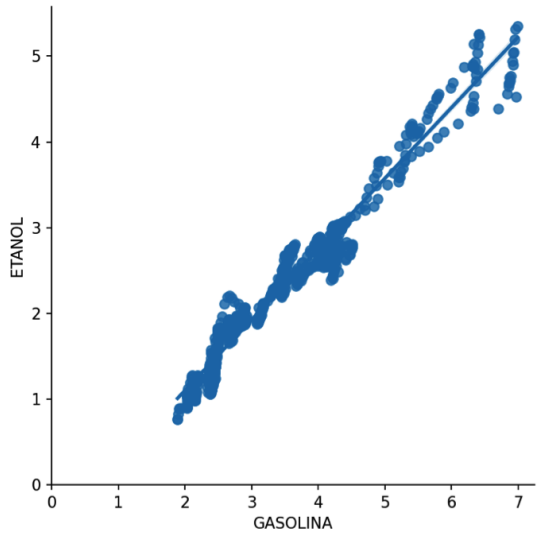


Figura 7: Gráfico de dispersão entre as séries históricas.

4 Dados e Resultados

4.1 Preparação dos Dados

4.1.1 Tratamento dos Dados

Assim como apontado na fonte dos dados, foi constatado que as séries históricas apresentavam 14 valores faltantes (*missings*) relativos a algumas semanas em que não foram realizadas as coletas de preços.

Após o tratamento, pelo método *ffill*, o conjunto de dados apresenta 970 observações com as características apresentadas pela Tabela 3.

Tabela 3: Estatísticas do conjunto de dados tratados.

	Etanol	Gasolina
n	970	970
Mínimo	0,766	1,888
1º Quartil	1,366	2,429
Média	2,124	3,244
Mediana	1,884	2,764
3º Quartil	2,622	3,967
Máximo	5,348	6,992
Desvio-padrão	0,919	1,091

Ressalta-se que não foram constatados casos de dados redundantes ou *outliers*.

4.1.2 Análise de Tendência e Sazonalidade das Séries Temporais

Realizada a decomposição dos dados a partir dos modelos aditivos e multiplicativos, conforme demonstrado nas Figuras 8-11, foram constatadas evidências de existência de tendência e sazonalidade nas séries de dados. Na decomposição de ambas as séries históricas, os valores residuais no modelo aditivo se apresentam de forma dispersa em torno do valor 0, enquanto no modelo multiplicativo os resíduos se apresentam de forma mais uniforme em torno do 1. Pelo comportamento menos disperso dos resíduos no modelo multiplicativo, esse modelo se demonstra mais razoável para treinamento.

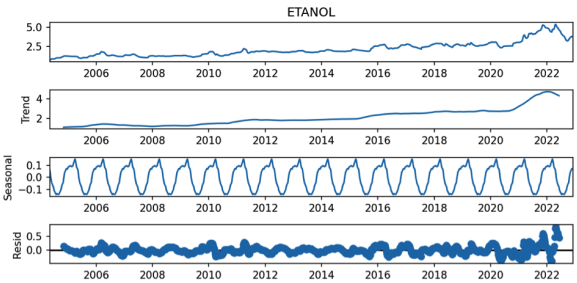


Figura 8: Decomposição da série do etanol pelo modelo aditivo.

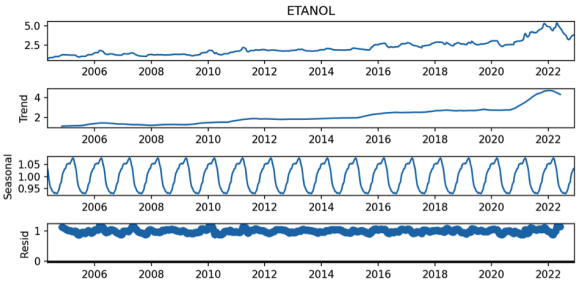


Figura 9: Decomposição da série do etanol pelo modelo multiplicativo.

4.1.3 *Data Split* da Série

Com o objetivo de desenvolver modelos de série temporal para predição dos valores futuros do preço de revenda do etanol, o conjunto de dados é dividido em três bases de dados: treino, teste e *deployment*. Enquanto a base de treino é utilizada para o treinamento dos modelos, a base de teste é utilizada para comparação e avaliação do modelo mais adequado. A base de *deployment* é utilizada para implementação da ferramenta desenvolvida,

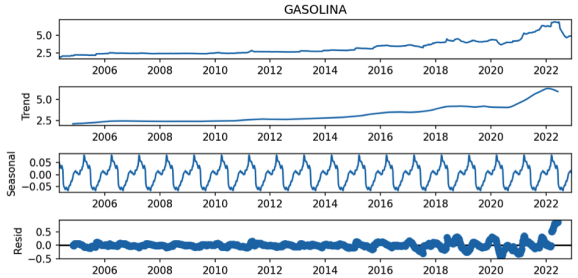


Figura 10: Decomposição da série da gasolina pelo modelo aditivo.

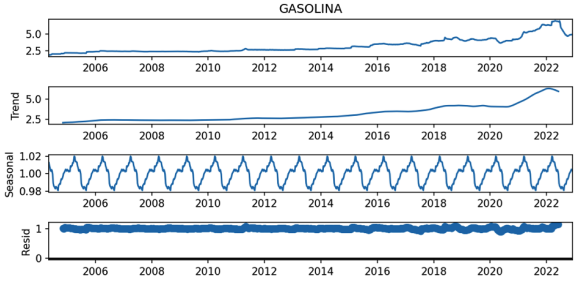


Figura 11: Decomposição da série da gasolina pelo modelo multiplicativo.

com uma simulação de um caso real.

Respeitando a estrutura temporal, a aleatoriedade não é aplicada na divisão dos dados. A base de treinamento consiste nas primeiras 70% observações (679 semanas). A base de teste consiste na 30% das observações finais (290 semanas), com exceção da última observação que é reservada à base de *deployment* (1 semana).

4.2 Treinamento (Modelagem)

Neste estudo, para desenvolvimento de um modelo de aprendizado de máquina para previsão dos preços do etanol com base numa série histórica de preços da gasolina, é proposto o treinamento com os modelos Prophet, LSTM e *Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors* (SARIMAX). No treinamento de todos os modelos, a série histórica com os preços da gasolina será adotada como hiperparâmetro.

4.2.1 Ambiente de Desenvolvimento

O desenvolvimento deste estudo é realizado no ambiente de execução colaborativo Google Colab, utilizando-se a linguagem Python 3.8.10. O conjunto de dados está armazenado em um arquivo texto no formato *Comma Separated Values* (CSV). A execução dos *notebooks* é realizada utilizando-se as bibliotecas: `Matplotlib`[2] 3.5.3 e `seaborn`[3] 0.11.2, para visualizações gráficas; `NumPy`[4] 1.22.4 e `pandas`[5] 1.3.5, para tratamento e manipulação dos dados; `pmdarima`[6] 1.7.1, para geração e treinamento do modelo SARIMAX; `Prophet`[7] 1.1.2, para geração e treinamento do modelo Prophet; `Sklearn`[8] 1.0.2, para avaliação das métricas; `Statsmodels`[9] 0.12.2 , para utilização de modelos estatísticos; e `TensorFlow`[10] 2.11.0, para geração e treinamento do modelo LSTM.

4.2.2 Definição do Método de Comparação e Avaliação dos Modelos

Conforme explicado por Faceli *et al.* (2021), na aplicação de algoritmos de aprendizado de máquina a problemas reais, em geral, o conhecimento é provido unicamente pelo conjunto de exemplos, a partir do qual a indução de um modelo preditivo/descritivo é então realizada. De maneira geral, pode-se afirmar que não existe técnica universal, ou seja, não é possível estabelecer *a priori* que uma técnica de aprendizado de máquina em particular se sairá melhor na resolução de qualquer tipo de problema. Ainda que um único algoritmo seja escolhido, pode ser necessário realizar ajustes

[2]THE MATPLOTLIB DEVELOPMENT TEAM. Matplotlib: Visualization with Python, 2023. Página inicial. Disponível em: `https://matplotlib.org/`. Acesso em: 25/02/2023.

[3]WASKOM, Michael. seaborn: statistical data visualization, 2022. Página inicial. Disponível em: `https://seaborn.pydata.org/`. Acesso em: 25/02/2023.

[4]NUMPY. NumPy, 2023. Página inicial. Disponível em: `https://numpy.org/`. Acesso em: 25/02/2023.

[5]PANDAS. pandas - Python Data Analysis Library, 2023. Página inicial. Disponível em: `https://pandas.pydata.org/`. Acesso em: 25/02/2023.

[6]PYTHON SOFTWARE FOUNDATION. pmdarima · PyPI, 2023. Página inicial. Disponível em: `https://pypi.org/project/pmdarima/`. Acesso em: 25/02/2023.

[7]FACEBOOK OPEN SOURCE. Prophet — Forecasting at scale., 2023. Página inicial. Disponível em: `https://facebook.github.io/prophet/`. Acesso em: 25/02/2023.

[8]SCIKIT-LEARN. scikit-learn: machine learning in Python, 2023. Página inicial. Disponível em: `https://scikit-learn.org/stable/`. Acesso em: 25/02/2023.

[9]PERKTOLD, JOSEF; SEABOLD, SKIPPER, TAYLOR, JONATHAN. Introduction – statsmodels, 2022. Página inicial. Disponível em: `https://www.statsmodels.org/`. Acesso em: 25/02/2023.

[10]TENSORFLOW. TensorFlow, 2023. Página inicial. Disponível em: `https://www.tensorflow.org/`. Acesso em: 25/02/2023.

em seus parâmetros livres, o que leva à obtenção de múltiplos modelos para os mesmos dados.

A validação de qualquer nova técnica de aprendizado de máquina proposta geralmente envolve a realização de experimentos controlados, em que se demonstre a sua efetividade na solução de diferentes problemas, representados por seus conjuntos de dados associados. É recomendável seguir procedimentos que garantam a correção, a validade e a reprodutibilidade dos experimentos realizados e, mais importante, das conclusões obtidas a partir de seus resultados.

Essa avaliação experimental de um algoritmo de aprendizado de máquina pode ser realizada segundo diferentes aspectos, tais como acurácia do modelo gerado, compreensibilidade do conhecimento extraído, tempo de aprendizado, requisitos de armazenamento do modelo, entre outros.

A avaliação de um modelo preditivo é normalmente realizada por meio da análise do desempenho do preditor gerado por ele na rotulação de novos objetos, não apresentados previamente em seu treinamento (Monard & Baranauskas, 2003).

No caso de problemas de regressão, o erro de hipótese pode ser calculado pela distância entre o valor y_i conhecido e aquele predito pelo modelo, $\hat{y}_i$.

Neste estudo, para seleção do modelo de treinamento mais adequado são adotadas as métricas *Root Mean Square Error* (RMSE), *Mean Absolute Percentage Error* (MAPE) e *Mean Absolute Error* (MAE), dadas, respectivamente, por:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}, \ \text{MAPE} = \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i}$$

e

$$\text{MAE} = \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{n}.$$

4.2.3 Treinamento pelo Modelo Prophet

O Prophet é um modelo de regressão aditiva com uma parte linear ou uma curva de tendência de crescimento logístico. Inclui um componente sazonal anual modelado usando série de Fourier e um componente sazonal semanal modelado usando variáveis fictícias (Amazon Web Services, 2023).

Disponível nas linguagens Python e R, é um *framework* (biblioteca) de código aberto, mantido e desenvolvido pelo Facebook. Originalmente, foi construído para resolver problemas específicos do Facebook envolvendo previsões em séries temporais. Em 2017, pesquisadores do Facebook publicaram o trabalho denominado "*Forecasting at Scale*" ("Previsões em Escala", em tradução livre), que introduziu o projeto de código aberto Facebook Prophet, oferecendo a analistas e cientistas de dados de todo o mundo uma ferramenta ágil, poderosa e acessível para modelagem de séries temporais (Krieger, 2021).

É otimizado para tarefas cujo conjunto de dados apresenta um longo período de observações históricas, fortes características de sazonalidades, irregulares e importantes eventos anteriormente conhecidos, pontos de dados faltantes ou de *outliers*, e tendências de crescimento não lineares que estão se aproximando de um ponto de limite ou de saturação.

Com sua abordagem *analyst-in-the-loop*, ao combinar previsões automáticas com previsões de analistas em *loop*, mesmo em suas configurações padrões, o Prophet é capaz de realizar previsões precisas, com menor esforço, mas sem limitar-se aos resultados de um procedimento completamente automático, caso a previsão não seja satisfatória. A Figura 12 ilustra a abordagem *analyst-in-the-loop*.

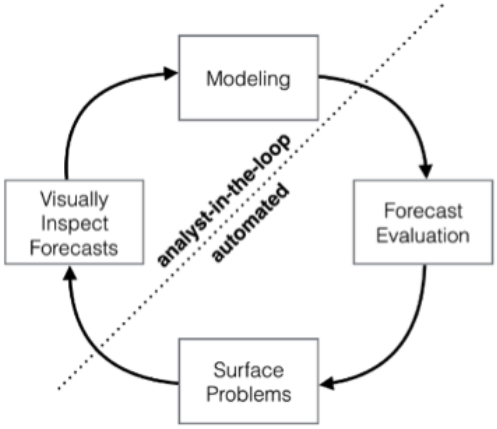


Figura 12: Abordagem *analyst-in-the-loop* do Prophet. Fonte: Letham & Taylor (2017).

O Prophet adota um modelo de séries temporais decomposto em três componentes princi-

pais: tendência não periódica, $g(t)$; sazonalidades periódicas, $s(t)$; feriados e datas importantes, $h(t)$; e erros, ϵ, que não são modelados pelas outras componentes (Letham & Taylor, 2017). A decomposição da série temporal, $y(t)$, pode ser representada pela seguinte equação:

$$y(t) = g(t) + s(t) + h(t) + \epsilon.$$

No desenvolvimento deste estudo, para treinamento do modelo Prophet, são consideradas as configurações padrões do *framework*, adotando-se apenas como parâmetro exógeno (hiperparâmetro) a série histórica de preços da gasolina.

Na Figura 13 é apresentada a representação gráfica das séries históricas do etanol e da gasolina com os valores previstos para a base de teste (30% das observações finais), obtidos após o treinamento com o modelo Prophet.

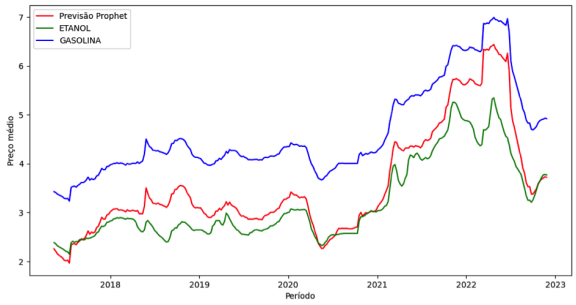


Figura 13: Comparação das séries com os valores previstos pelo modelo Prophet.

4.2.4 Treinamento pelo Modelo LSTM

O LSTM é uma arquitetura de rede neural recorrente (RNN) que memoriza valores ao longo de intervalos arbitrários. As redes LSTM são eficazes para classificar, processar e prever séries temporais com desconhecidos intervalos de tempo (*time lags*). A relativa insensibilidade à extensão das lacunas (*gap*) dá uma vantagem à LSTM sobre versões alternativas de RNNs, modelos ocultos de Markov e outros métodos sequenciais de aprendizado (Kang, 2017).

Redes LSTM foram especialmente desenvolvidas para superar o problema da dependência de longo prazo enfrentado pelas RNNs, em função do problema da dissipação do gradiente (*vanishing gradient problem*). LSTMs possuem conexões recorrentes (*feedback*) que as diferenciam das mais tradicionais redes neurais diretas (*feedforward*). Essa propriedade possibilita às LSTMs processar sequências inteiras de dados (séries temporais, por exemplo) sem tratar cada dado da sequência de forma independente, mas sim, retendo informação útil sobre dados anteriores da sequência para auxiliar no processamento dos novos dados. Como resultado, LSTMs são particularmente boas no processamento de sequência de dados como texto, voz e séries temporais em geral (Dolphin, 2020).

De forma básica, para explicar seu funcionamento, a saída de uma rede LSTM em um particular ponto no tempo é dependente de três fatores:

- a atual memória de longo prazo da rede – conhecida por *cell state*;
- a saída obtida no ponto previamente anterior – conhecida por *previous hidden state*; e
- a atual informação de entrada da série temporal – *input data* x_t (Dolphin, 2020).

Na Figura 14 é ilustrado um diagrama de uma célula LSTM.

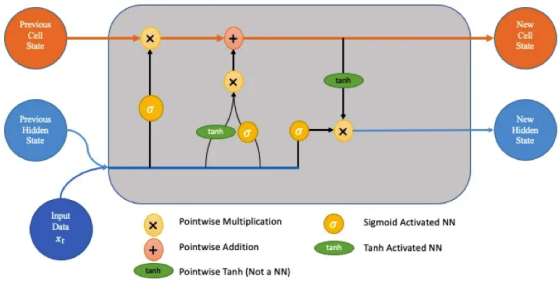


Figura 14: Diagrama de uma célula LSTM. Fonte: Dolphin (2020).

Para controle do *input*, da memorização e do *output* da informação contida numa sequência de dados, redes LSTM típicas possuem 3 *gates*. Esses *gates* filtram as informações consideradas importantes e cada um constitui sua própria rede neural. Em ordem, são os seguintes:

i) *forget gate*, que decide quais partes do *cell state* continuam importantes;

ii) *remember gate*, que decide quais informações da memória de curto prazo devem ser adicionadas ao *cell state*; e

iii) *output gate*, que decide quais partes do *cell state* são importantes no instante atual para

gerar o *output* (Matsumoto *et al.*, 2019).

Na Figura 15 é ilustrado um esquema completo de uma rede LSTM.

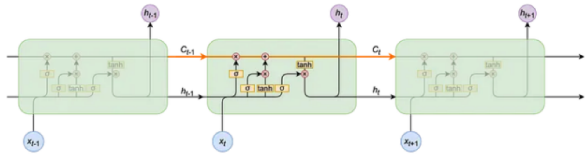


Figura 15: Esquema de uma rede LSTM. Fonte: Matsumoto *et al.* (2019).

No desenvolvimento deste estudo, para treinamento do modelo LSTM, é adotada uma rede com cinco camadas ocultas com otimizador Adam. Assim como no modelo anterior, a série histórica de preços da gasolina é adotada como parâmetro exógeno (hiperparâmetro).

Na Figura 16 é apresentada a representação gráfica das séries históricas do etanol e da gasolina com os valores previstos para a base de teste (30% das observações finais), obtidos após o treinamento com o modelo LSTM.

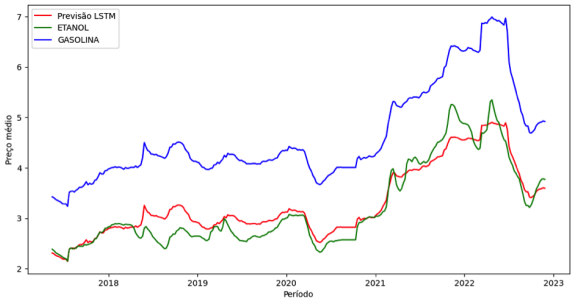


Figura 16: Comparação das séries com os valores previstos pelo modelo LSTM.

4.2.5 Treinamento pelo Modelo SARIMAX

O SARIMAX é um modelo ARIMA com sazonalidade e adição de uma ou mais variáveis exógenas.

Modelos autorregressivos são ferramentas clássicas para análise de séries temporais. O modelo ARIMA é uma classe de modelos lineares que utiliza valores históricos (autorregressão, médias móveis) para prever valores futuros. ARIMA significa *Autoregressive Integrated Moving Average* e cada uma das técnicas contribui para a previsão final. O modelo SARIMA, similarmente, utiliza valores históricos, mas também inclui a contribuição da sazonalidade para a previsão. SARIMA significa *Seasonal* ARIMA. A importância da sazonalidade é bastante evidente à medida que o ARIMA falha em encapsular essa informação implicitamente (Bajaj, 2023).

AR, MA, ARMA, ARIMA e ARIMAX são igualmente modelos lineares para análise de séries temporais. Considerados como casos especiais do modelo SARIMAX, a previsão de valores futuros da série pode ser compreendida através da definição desses seus modelos precursores (GitHub, 2023).

i. Modelos autorregressivos (AR)

Em um modelo de autorregressão, a variável de interesse é prevista usando uma combinação linear de valores passados dessa variável. O termo autorregressão indica que é uma regressão da variável contra ela mesma. Ou seja, usa-se valores defasados da variável predita como valores de entrada para prever valores futuros (Bajaj, 2023). A componente autorregressiva é representada por AR(p), com o parâmetro p determinando o grau da defasagem (Artley, 2022).

Dada uma série temporal y_t, um modelo AR(p) é matematicamente especificado por:

$$y_t = \beta + \epsilon_t + \sum_{i=1}^{p} \theta_i y_{t-i},$$

sendo p o número de intervalos de regressão, ϵ_t o ruído no tempo t e β uma constante.

Esta equação pode ser escrita de forma mais concisa através do uso do operador *lag*, L:

$$L^n y_t = y_{t-n}.$$

Sendo $\Theta(L)^p$ uma função polinomial (ordem p) de L, um modelo autorregressivo é definido por:

$$y_t = \Theta(L)^p y_t + \epsilon_t,$$

sendo a constante β adicionada ao polinômio Θ.

ii. Modelos de médias móveis (MA)

Os modelos de médias móveis usam erros de previsão passados, em vez de valores passados, em um modelo de regressão para prever valores futuros

(Bajaj, 2023). A componente de médias móveis é representada por MA(q), em que q é o grau de defasagem dos erros na previsão.

Um modelo MA(q) pode ser especificado por:

$$y_t = \Phi(L)^q \epsilon_t + \epsilon_t,$$

sendo q o número de intervalos de regressão de erro e Φ definido analogamente a Θ.

iii. Modelos autorregressivos de médias móveis (ARMA)

Os modelos autorregressivos de médias móveis, ARMA(p, q), são simplesmente uma soma dos modelos AR(p) e MA(q):

$$y_t = \Theta(L)^p y_t + \Phi(L)^q \epsilon_t + \epsilon_t.$$

iv. Modelos autorregressivos integrados de médias móveis (ARIMA)

Um modelo possuir uma componente integrada representa qualquer diferenciação que deve ser aplicada para tornar os dados estacionários (Bajaj, 2023).

A componente de diferenciação sazonal é representada por I(d), em que o parâmetro d representa o número de transformações (diferenciações) necessárias para tornar a série estacionária (Artley, 2022).

Para ajudar a lidar com dados não estacionários, introduz-se um operador de integração Δ^d, definido da seguinte forma:

$$y_t^{[d]} = \Delta^d y_t = y_t^{[d-1]} - y_{t-1}^{[d-1]},$$

sendo $y_t^{[0]} = y_t$ e d a ordem de diferenciação utilizada.

Ajustando o modelo ARMA(p, q) para $y_t^{[d]}$, ao invés de y_t, tem-se:

$$y_t^{[d]} = \Theta(L)^p y_t^{[d]} + \Phi(L)^q \epsilon_t^{[d]} + \epsilon_t^{[d]},$$

sendo equivalente ao modelo ARIMA(p, q, d) em y_t:

$$\Delta^d y_t = \Theta(L)^p \Delta^d y_t + \Phi(L)^q \Delta^d \epsilon_t + \Delta^d \epsilon_t.$$

Reorganizando a equação e adicionando as constantes aos polinômios Θ e Φ, obtém-se:

$$\Theta(L)^p \Delta^d y_t = \Phi(L)^q \Delta^d \epsilon_t.$$

v. Modelos autorregressivos integrados de médias móveis sazonais (SARIMA)

Os modelos SARIMA levam em consideração a sazonalidade, aplicando essencialmente a um modelo ARIMA *lags* que são múltiplos inteiros da sazonalidade. Uma vez que a sazonalidade é modelada, um modelo ARIMA é aplicado ao restante para capturar a estrutura não sazonal (GitHub, 2023).

O modelo SARIMA é muito semelhante ao modelo ARIMA, exceto que há um conjunto adicional de componentes autorregressivos e de médias móveis. Os atrasos adicionais são compensados pela frequência da sazonalidade (Artley, 2022).

As componentes Autorregressiva – AR(p), Integrada – I(d) e Média Móvel – MA(q) do modelo permanecem como as do ARIMA. A adição de sazonalidade adiciona robustez ao modelo SARIMA (Bajaj, 2023).

O modelo SARIMA é representado como: SARIMA$(p, d, q) \times (P, D, Q)m$, em que (p, d, q) representa as componentes não sazonais e $(P, D, Q)m$ as componentes sazonais.

Os modelos SARIMA permitem diferenciar os dados pela frequência sazonal e também pela diferenciação não sazonal. Saber quais parâmetros são os melhores pode ser facilitado com o emprego de *frameworks* para pesquisa de parâmetros, como o `pmdarima` (Artley, 2022).

Para uma melhor compreensão do modelo e de suas componentes (não sazonais e sazonais), suponha uma série temporal y_t com sazonalidade m. A sazonalidade pode ser eliminada por diferenciação, aplicando-se o operador diferencial Δ_m^D para tomar as diferenças sazonais da série temporal. O parâmetro m representa o intervalo de tempo (número de *lags*) compreendido por um período completo de sazonalidade. O parâmetro D assume um significado semelhante ao parâmetro d dos modelos ARIMA, mas sendo considerados os *lags* sazonais.

Usando-se então os *lags* sazonais, qualquer estrutura restante pode ser capturada aplicando-se um modelo ARMA(P, Q) para os valores diferenciados (GitHub, 2023):

$$\Delta_m^D y_t = \theta(L^m)^P \Delta_m^D y_t + \phi(L^m)^Q \Delta_m^D \epsilon_t + \Delta_m^D \epsilon_t,$$

sendo L^m no lugar de L, P e Q os operadores de *lag* sazonais.

Assim como no modelo ARIMA, manipular a equação e adicionar as constantes aos polinômios produz a seguinte forma concisa:

$$\theta(L^m)^P \Delta_m^D y_t = \phi(L^m)^Q \Delta_m^D \epsilon_t.$$

Com qualquer sazonalidade agora removida, é possível aplicar outro modelo ARIMA(p, d, q) a $\Delta_m^D y_t$, multiplicando o modelo sazonal pelo novo modelo ARIMA.

Essa é, então, a fórmula geral do modelo SARIMA$(p, d, q) \times (P, D, Q)m$:

$$\Theta(L)^p \theta(L^m)^P \Delta^d \Delta_m^D y_t = \Phi(L)^q \phi(L^m)^Q \Delta^d \Delta_m^D \epsilon_t.$$

vi. Modelos autorregressivos integrados de médias móveis sazonais com variáveis exógenas (SARIMAX)

Os modelos ARIMAX e SARIMAX simplesmente levam em consideração variáveis exógenas, ou seja, variáveis medidas no intervalo t que influenciam o valor da série temporal no intervalo t, mas que não são também autorregredidas (GitHub, 2023).

Para n variáveis exógenas definidas em um intervalo t, denotadas por x_t^i, com $i \leq n$, e coeficientes β_i, o modelo ARIMAX(p, d, q) é matematicamente especificado por:

$$\Theta(L)^p \Delta^d y_t = \Phi(L)^q \Delta^d \epsilon_t + \sum_{i=1}^{n} \beta_i x_t^i$$

e o modelo SARIMAX$(p, d, q) \times (P, D, Q)m$ é especificado por:

$$\Theta(L)^p \theta(L^m)^P \Delta^d \Delta_m^D y_t = \Phi(L)^q \phi(L^m)^Q \Delta^d \Delta_m^D \epsilon_t + \sum_{i=1}^{n} \beta_i x_t^i.$$

É interessante pensar que, tecnicamente, todos os fatores exógenos ainda são indiretamente modelados na previsão do modelo histórico. Ainda assim, se incluídos dados externos, o modelo responderá muito mais rapidamente ao seu efeito do que apenas contando com a influência de termos atrasados (Artley, 2022).

No desenvolvimento deste estudo, para treinamento do modelo SARIMAX, assim como nos modelos anteriores, a série histórica de preços da gasolina é adotada como parâmetro exógeno (hiperparâmetro).

Na Figura 17 é apresentada a representação gráfica das séries históricas do etanol e da gasolina com os valores previstos para a base de teste (30% das observações finais), obtidos após o treinamento com o modelo SARIMAX.

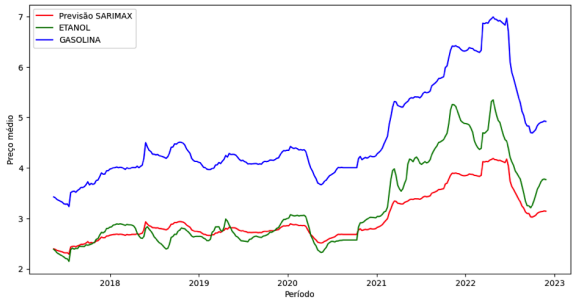


Figura 17: Comparação das séries com os valores previstos pelo modelo SARIMAX.

4.3 Avaliação

Todas as métricas adotadas para seleção do modelo de treinamento mais adequado (RMSE, MAPE e MAE), apontam o modelo LSTM como melhor escolha, conforme observado na Tabela 4.

Tabela 4: Métricas para avaliação dos modelos.

Modelo	RMSE	MAPE	MAE
SARIMAX	0,44	0,08	0,31
LSTM	0,25	0,06	0,20
Prophet	0,52	0,11	0,38

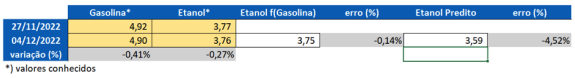
4.4 *Deployment*

Para a implementação da ferramenta desenvolvida, é imaginada uma situação em que se deseja prever o novo valor de revenda do etanol, ao ser divulgado um reajuste no preço da gasolina (combustível com valor de saída das refinarias controlado pela Petrobras).

Para a análise da efetividade da ferramenta, é adotada como base de dados a última semana da série histórica dos preços médios considerada.

Conforme observado na Figura 18, conhecidos os valores para as semanas de 27/11/2022 e 04/12/2022, é sabida a queda real nos preços da gasolina e do etanol para R$ 4,90 e R$ 3,76, respectivamente. Percentualmente, uma queda de 0,41% e 0,27%.

Com base na variação de preço da gasolina, o preço



	Gasolina*	Etanol*	Etanol f(Gasolina)	erro (%)	Etanol Predito	erro (%)
27/11/2022	4,92	3,77				
04/12/2022	4,90	3,76	3,75	-0,14%	3,59	-4,52%
variação (%)	-0,41%	-0,27%				

*) valores conhecidos

Figura 18: Resultados do *deployment*.

estimado do etanol para a semana de 04/12/2022 é de R$ 3,75. Um erro percentual de 0,14%.

Caso adotado o modelo LSTM para previsão dos preços do etanol, é previsto o valor de aproximadamente R$ 3,59. Um erro percentual de 4,52%.

5 Conclusão

Os preços de revenda do etanol combustível claramente sofrem uma forte influência dos preços da gasolina. Embora mais barato, o menor poder calorífico do etanol combustível não oferece vantagem competitiva suficiente para que o seu preço seja muito superior a 70% do preço da gasolina. No estudo realizado, o limite de 70% foi superado somente em 15% das observações analisadas.

O forte grau de associação linear positiva entre as séries históricas observadas evidencia que os produtores de etanol combustível adotam, de fato, como referência o preço da gasolina. Apesar desse último ser um produto derivado de uma das *commodities* mais afetadas (tanto nacionalmente como internacionalmente) por questões políticas e econômicas, ambas as séries históricas apresentam um comportamento de distribuição estatístico muito similar. Uma análise superficial de ambas as séries históricas sugere que os produtores de etanol, quando possível, tendem a acompanhar a variação de preço da gasolina para maximizar sua margem de lucro.

Por exemplo, com a publicação da Lei Complementar nº 194, de 23 de junho de 2022, no estado de São Paulo a alíquota do Imposto sobre Circulação de Mercadorias e Serviços (ICMS) sobre as operações com gasolina foi reduzida de 25% para 18%, enquanto sobre as operações com etanol foi mantida a alíquota de 13,3%. Apesar disso, no período observado não há qualquer mudança significativa no comportamento de ambas as séries históricas. Segundo a agência de notícias epbr[11], a diminuição dos tributos acabou beneficiando o mercado de gasolina e, com isso, o preço do etanol enfrentou queda de 13%, fazendo com que a paridade entre os produtos, em julho, ficasse, na média, em 68,9%.

No desenvolvimento do modelo de aprendizado de máquina para previsão dos preços do etanol combustível, apesar dos melhores resultados obtidos com a rede LSTM, a adoção da série histórica da gasolina como variável exógena proporcionou aos três modelos sugeridos resultados com métricas consideradas igualmente boas. A gasolina demonstrou-se como um confiável hiperparâmetro para previsão dos preços do etanol.

Como trabalhos futuros, dado os resultados obtidos com a adoção da gasolina como hiperparâmetro, primeiramente propõe-se a ampliação do número de variáveis exógenas adotadas para treinamento dos modelos. É esperado que, com a adoção de hiperparâmetros como a taxa de câmbio e a cotação de algumas importantes *commodities* (petróleo e açúcar, por exemplo), se obtenha uma previsão mais realista para os preços do etanol combustível.

Visto não ter sido o principal objetivo do estudo, outra proposta de aprimoramento cabe à continuidade no desenvolvimento dos modelos de aprendizado de máquina apresentados e outros modelos atualmente em estudo e desenvolvimento pela comunidade científica. Conforme apontado nas seções iniciais, modelos como de regressão linear múltipla ou de decomposição temporal com tendência linear e sazonalidade poderão ser capazes de oferecer predições mais confiáveis.

Dado ainda os bons resultados com as redes LSTM, especialmente se propõe o desenvolvimento de estudos com modelos híbridos como, por exemplo, o proposto por Yan *et al.* (2019), em que redes neurais LSTM foram combinadas com técnicas *Stationary Wavelet Transform* (SWT).

Por fim, particularmente em função dos trabalhos desenvolvidos pela Secretaria da Fazenda e Planejamento do Estado de São Paulo para concessão de benefícios fiscais, espera-se que a ferramenta desenvolvida seja capaz de auxiliar na predição do comportamento do preço de determinado produto, baseada na adoção de produtos concorrentes ou similares.

[11] AGÊNCIA EPBR, 2020. Mercado do etanol sofre baixa e gasolina ganha espaço com ICMS menor. Disponível em: `https://epbr.com.br/mercado-do-etanol-sofre-baixa-e-gasolina-ganha-espaco-com-icms-menor/`. Acesso em: 06/03/2023.

Referências

[1] ALLEN, Genevera. AAAS: machine learning 'causing science crisis. *AITopics*, 2019. Disponível em: https://aitopics.org/doc/news:701AE8C7/. Acesso em: 26/06/2022.

[2] ALMEIDA, Edilberto Tiago de; JUSTO, Wellington Ribeiro; OLIVEIRA, Monaliza Ferreira de; SILVA, Carla Calixto da. Uma análise da demanda por combustíveis através do modelo almost ideal demand system para Pernambuco. *Revista de Economia e Sociologia Rural*, v.54, p.691-708, 2016.

[3] AMAZON Web Services, Inc. Amazon Forecast: Guia do desenvolvedor. AWS, 2023. Disponível em: https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-recipe-prophet.html. Acesso em: 14/02/2023.

[4] ANIFOWOSE, Fatai Adesina; LABADIN, Jane; ABDULRAHEEM, Abdulazeez. Ensemble machine learning: An untapped modeling paradigm for petroleum reservoir characterization. *Journal of Petroleum Science and Engineering*, v.151, p.480-487, 2017.

[5] ARTLEY, Brendan. Time Series Forecasting with ARIMA, SARIMA and SARIMAX. Towards Data Science, 2022. Disponível em: https://towardsdatascience.com/time-series-forecasting-with-arima-sarima-and-sarimax-ee61099e78f6. Acesso em: 20/02/2023.

[6] AZADEH, Aghaderi; GHADERI, S. F.; SOHRABKHANI, Sara. Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors. *Energy Conversion and management*, v.49, n.8, p.2272-2278, 2008.

[7] BAJAJ, Aayush. ARIMA & SARIMA: Real-World Time Series Forecasting. Neptune Labs, 2022. Disponível em: https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide. Acesso em: 02/03/2023.

[8] BEYCA, Omer Faruk; ERVURAL, Beyzanur Cayir; TATOGLU, Ekrem; OZUYAR, Pinar Gokcin; ZAIM, Selim. Using machine learning tools for forecasting natural gas consumption in the province of Istanbul. *Energy Economics*, v.80, p.937-949, 2019.

[9] DOLPHIN, Rian. LSTM Networks — A Detailed Explanation. Towards Data Science, 2020. Disponível em: https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9. Acesso em: 16/02/2023.

[10] FACELI, Katti; LORENA, Ana Carolina; GAMA, João; ALMEIDA, Tiago Agostinho de; CARVALHO, André C. P. L. F. de. Inteligência artificial: uma abordagem de aprendizado de máquina, Rio de Janeiro: LTC., 2.ed., p.148-150, 2021.

[11] GHODDUSI, Hamed; CREAMER, Germán G.; RAFIZADEH, Nima. Machine learning in energy economics and finance: A review. *Energy Economics*, v.81, p.709-727, 2019.

[12] GITHUB. From AR to SARIMAX: Mathematical Definitions of Time Series Models. GitHub Pages, 2023. Disponível em: https://phosgene89.github.io/sarima.html. Acesso em: 03/03/2023.

[13] GUPTA, Sudha; KAMBLI, Ruta; WAGH, Sushama; KAZI, Faruk. Support-vector-machine-based proactive cascade prediction in smart grid using probabilistic framework. *IEEE Transactions on Industrial Electronics*, v.62, n.4, p.2478-2486, 2014.

[14] HEINERMANN, Justin; KRAMER, Oliver. Machine learning ensembles for wind power prediction. *Renewable Energy*, v.89, p.671-679, 2016.

[15] JURADO, Sergio; NEBOT, Àngela; MUGICA, Fransisco; AVELLANA, Narcís. Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques. *Energy*, v.86, p.276-291, 2015.

[16] KANG, Eugine. Long Short-Term Memory (LSTM): Concept. Medium, 2017. Disponível em: https://medium.com/@kangeugine/long-short-term-memory-lstm-concept-cb3283934359. Acesso em: 16/02/2023.

[17] KRIEGER, Mitchell. Time Series Analysis with Facebook Prophet: How it works and How to use it. Towards

Data Science, 2021. Disponível em: https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a. Acesso em: 25/02/2023.

[18] LETHAM, Ben; TAYLOR, Sean J. Prophet: forecasting at scale. Meta, 2017. Disponível em: https://research.facebook.com/blog/2017/2/prophet-forecasting-at-scale/. Acesso em: 14/02/2023.

[19] LI, Zheng; ZHOU, Bo; HENSHER, David A. Forecasting automobile gasoline demand in Australia using machine learning-based regression. *Energy*, v.239, p.122312, 2022.

[20] MARUGÁN, Alberto Pliego; MÁRQUEZ, Fausto Pedro García; PEREZ, Jesus María Pinar; RUIZ-HERNÁNDEZ, Diego. A survey of artificial neural network in wind energy systems. *Applied Energy*, v.228, p.1822-1836, 2018.

[21] MATSUMOTO, Fernando; DUARTE, Guilherme; MURAKAMI, Leonardo. Redes Neurais — LSTM. Medium, 2019. Disponível em: https://medium.com/turing-talks/turing-talks-27-modelos-de-predi%C3%A7%C3%A3o-lstm-df85d87ad210. Acesso em: 17/02/2023.

[22] MONARD, Maria Carolina, BARANAUSKAS, José Augusto. Conceitos Sobre Aprendizado de Máquina. Sistemas Inteligentes Fundamentos e Aplicações. 1.ed. Barueri SP: Manole Ltda, 2003. p.89–114.

[23] SANTOS, Fábio Gouveia dos. Impacto da precificação dinâmica na revenda varejista de combustíveis. FGV, Tese de Doutorado, 2021.

[24] SCHUTTE, Giorgio Romano; BARROS, Pedro Silva. A geopolítica do etanol. IPEA, 2010.

[25] VOYANT, Cyril; NOTTON, Gilles; KALOGIROU, Soteris; NIVET, Marie-Laure; PAOLI, Christophe; MOTTE, Fabrice; FOUILLOY, Alexis. Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, v.105, p.569-582, 2017.

[26] YAN, Ke; LI, Wei; JI, Zhiwei; QI, Meng; DU, Yang. A hybrid LSTM neural network for energy consumption forecasting of individual households. *IEEE Access*, v.7, p.157633 157642, 2019.

[27] YU, Lean; MA, Yueming. A Data-Trait-Driven Rolling Decomposition-Ensemble Model for Gasoline Consumption Forecasting. *Energies*, v.14, n.15, p.4604, 2021.

[28] YU, Lean; MA, Yueming; MA, Mengyao. An effective rolling decomposition-ensemble model for gasoline consumption forecasting. *Energy*, v.222, p.119869, 2021.

[29] ZAIDI, Syed Mohammed Arshad; CHANDOLA, Varun; ALLEN, Melissa R.; SANYAL, Jibonananda; STEWART, Robert N.; BHADURI, Budhendra L.; MCMANAMAY, Ryan A. Machine learning for energy-water nexus: challenges and opportunities. *Big Earth Data*, v.2, n.3, p.228-267, 2018.

[30] ZENDEHBOUDI, Sohrab; REZAEI, Nima; LOHI, Ali. Applications of hybrid models in chemical, petroleum, and energy systems: A systematic review. *Applied Energy*, v.228, p.2539-2566, 2018.

[31] ZENG, Yuyun; LIU, Jingquan; SUN, Kaichao; HU, Lin-wen. Machine learning based system performance prediction model for reactor control. *Annals of Nuclear Energy*, v.113, p.270-278, 2018.

[32] ZHANG, Jindai; ZHAO, Jinlou. Trend-and Periodicity-Trait-Driven Gasoline Demand Forecasting. *Energies*, v.15, n.10, p.3553, 2022.