CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

UTILIZAÇÃO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL NA PREDIÇÃO MENSAL DA PRODUÇÃO BRASILEIRA DE ETANOL COMBUSTÍVEL

Antonio Carlos Silva Mello

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Produção e Sistemas, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Eduardo Soares Ogasawara
Diego Moreira de Araújo Carvalho

Rio de Janeiro,
Setembro de 2024

Utilização de modelos de inteligência artificial na predição mensal da produção brasileira de etanol combustível

Dissertação de Mestrado em Engenharia de Produção e Sistemas, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Antonio Carlos Silva Mello

Aprovada por:

Presidente, Prof. Eduardo Soares Ogasawara, D.Sc. (orientador)

Diego Moreira de Araujo Carvalho, D.Sc. (coorientador)

Rafael Garcia Barbastefano, D.Sc.

Marcelle Candido Cordeiro, D.Sc.

Rio de Janeiro,

Setembro de 2024

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

M527 Mello, Antonio Carlos Silva
Utilização de modelos de inteligência artificial na predição mensal da produção brasileira de etanol combustível / Antonio Carlos Silva Mello. — 2024.
41f. : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2024.
Bibliografia : f. 36-41
Orientador: Eduardo Soares Ogasawara
Coorientador: Diego Moreira de Araújo Carvalho

1. Inteligência artificial. 2. Aprendizado do computador. 3. Álcool como combustível. 4. Análise de séries temporais. I. Ogasawara, Eduardo Soares. (Orient.). II. Carvalho, Diego Moreira de Araújo. (Coorient.). III. Título.

CDD 006.3

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

DEDICATÓRIA

Dedico esta dissertação à minha esposa, Sandra Valéria, e aos meus filhos, Eduardo e Luciano, pela paciência e apoio ao longo de toda esta jornada. Também dedico ao meu querido pai, José Vieira, que acompanhou boa parte dessa caminhada e cuja memória me inspirou a continuar, mesmo na sua ausência.

RESUMO

Utilização de modelos de inteligência artificial na predição mensal da produção brasileira de etanol combustível

Antonio Carlos Silva Mello

Orientadores:
Eduardo Soares Ogasawara
Diego Moreira de Araújo Carvalho

Resumo da Dissertação submetida ao Programa de Pós-graduação em Engenharia de Produção e Sistemas do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do título de mestre.

O etanol é um dos principais biocombustíveis produzidos no Brasil. Em 2023, dados da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) indicaram que o volume total de etanol combustível vendido no Brasil foi de pouco mais de 28 milhões de metros cúbicos (m^3), o que correspondeu a quase 22% do volume total de combustíveis líquidos vendidos no país. Dois tipos de etanol são produzidos no Brasil: i) o etanol hidratado, utilizado diretamente como combustível veicular, e ii) o etanol anidro, que é adicionado atualmente em uma proporção de 27 % na composição da gasolina comum. Apenas seis estados são responsáveis por aproximadamente 90% da produção brasileira de etanol. O desafio logístico envolvido decorre tanto da sazonalidade da produção quanto da necessidade de transportar o etanol dos pontos de produção até as redes de distribuição e revenda. Considerando os recentes avanços em Inteligência Artificial, Esse desafio levanta as seguintes questões de pesquisa: Podemos melhorar a previsão mensal da produção de etanol hidratado e anidro para os principais estados produtores brasileiros usando Modelos de Inteligência Artificial? Como os dados devem ser preparados para tal abordagem? Este estudo busca contribuir com o planejamento logístico correlato, ao empregar uma metodologia de Inteligência Artificial (IA) centrada em dados (*Data-Centric AI*), para auxiliar no desenvolvimento e seleção de modelos de IA para séries temporais de produção de etanol nos principais estados produtores. Os resultados experimentais quantitativos demonstram que a metodologia proposta supera significativamente os modelos de referência em precisão de previsão para a produção de etanol nos cenários avaliados.

Palavras-chave: Data-centric AI, Aprendizado de Máquina, Modelos Auto-Regressivos, Etanol combustível, Modelos de Inteligência Artificial.

Rio de Janeiro,

Setembro de 2024

ABSTRACT

Utilização de modelos de inteligência artificial na predição mensal da produção brasileira de etanol combustível

Antonio Carlos Silva Mello

Advisors:
Eduardo Soares Ogasawara
Diego Moreira de Araújo Carvalho

Abstract of dissertation submitted to Programa de Pós-graduação em Engenharia de Produção e Sistemas - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Ethanol is one of the main biofuels produced in Brazil. In 2023, data from the National Agency of Petroleum, Natural Gas, and Biofuels (ANP) indicated that the total volume of fuel ethanol sold in Brazil was just over 28 million cubic meters (m^3), corresponding to almost 22% of the total volume of liquid fuels sold in the country. Two types of ethanol are produced in the Brazil: i) hydrous ethanol, used directly as vehicle fuel, and ii) anhydrous ethanol, which is currently added at a proportion of 27% to regular gasoline. Only six states are responsible for approximately 90% of Brazilian ethanol production. The logistical challenge involved stems from both the seasonality of production and the need to transport ethanol from production points to distribution and retail networks. Considering recent advances in Artificial Intelligence, this challenge raises the following research questions: Can we improve the monthly forecast of hydrated and anhydrous ethanol production for the main producing states in Brazil using Artificial Intelligence Models? How should the data be prepared for such an approach? This study aims to contribute to related logistical planning by employing a Data-centric Artificial Intelligence (AI) methodology to assist in the development and selection of AI models for time series of ethanol production in the main producing states. Quantitative experimental results demonstrate that the proposed methodology significantly outperforms benchmark models in forecast accuracy for ethanol production in the evaluated scenarios.

Key-words: Data-centric AI, Machine Learning, Auto-Regressive Models, Fuel ethanol, Artificial Intelligence Models.

Rio de Janeiro,

Setembro de 2024

Sumário

I Introdução **1**

II Referencial teórico **3**

II.1 Produção brasileira de etanol 3

II.2 Séries temporais 6

II.3 Modelos Auto-Regressivos 7

II.4 Modelos ARIMA 8

II.5 Data-centric AI (DAI) 9

II.6 AIM 10

II.6.1 MLP e ELM 11

II.6.2 Modelos de Aprendizado Profundo (LSTM e Conv1D) 13

II.6.3 Regressão de Suporte Vetorial (SVR) 14

II.6.4 Modelos baseados em incerteza 15

II.7 Otimização de hiperparâmetros 16

III Trabalhos relacionados **17**

III.1 Predição da produção de etanol ou outros biocombustíveis 19

III.2 Predição de consumo de gasolina e/ou etanol combustível 20

III.3 Discussão sobre os trabalhos relacionados 23

IV Metodologia **26**

V Avaliação experimental **29**

V.1 Conjuntos de dados 29

V.2 Implementação 29

V.3 Resultados 31

VI Conclusões **35**

Referências **36**

Lista de Figuras

II.1 Produção mensal brasileira de etanol anidro e hidratado (m^3), entre os anos de 2012 e 2023. Fonte: Adaptado de ANP [2024a] 3

II.2 Utilização de cana de açúcar x milho na produção brasileira de etanol. Fonte: Adaptado de UNICAdata [2023] 5

II.3 Estados brasileiros que utilizam o milho como matéria prima para a produção de etanol - Comparação dos volumes produzidos. Fonte: Adaptado de UNICAdata [2023] 6

II.4 Decomposição de uma série temporal em seus componentes 7

II.5 Funções de ativação sigmoide, ReLU e softplus 12

II.6 Exemplo de MLP com duas entradas X_1 e X_2, uma camada oculta com dois neurônios a_1 e a_2 e um neurônio de saída y_1 12

II.7 Exemplo de Regressão de Suporte Vetorial 15

III.1 Aplicação do fluxograma PRISMA para a seleção das referências a partir das bases [Page et al., 2021] 18

IV.1 Processo D-AI^2-M - Seleção dos melhores modelos para previsão mensal brasileira da produção de etanol) 27

V.1 Séries temporais de produção de etanol anidro e hidratado dos principais estados produtores. Fonte: Adaptado de ANP [2024a] 30

V.2 Detalhes da validação cruzada de séries temporais (avaliação com origem móvel) 31

Lista de Tabelas

II.1 Produção e consumo de etanol por estado brasileiro entre os anos de 2019 e 2023 (dados da ANP [ANP, 2024a,b]) 4

II.2 Mudanças nas regras para a mistura de etanol anidro na gasolina comum brasileira 5

III.1 Consultas realizadas nas bases Scopus e Web of Science 17

III.2 Trabalhos relacionados com a pesquisa proposta 25

V.1 Hiperparâmetros para a otimização dos D-AI^2-M 30

V.2 Resultados da comparação entre ARIMA e os AIM (R^2 médio na validação cruzada de série temporal) 32

V.3 Resultados da comparação entre ARIMA e os DAI (R^2 médio na validação cruzada de séries temporais) 32

V.4 Resultados da comparação Local entre ARIMA e os D-AI^2-M (R^2 médio na validação cruzada de séries temporais) 33

V.5 Resultados da comparação Global entre ARIMA e os D-AI^2-M (R^2 médio na avaliação e na validação cruzada de séries temporais) 34

Lista de Acrônimos

AIM	Artificial Inteligence Model.
ANP	Agência Nacional de Petróleo, Gás Natural e Biocombustíveis.
ARIMA	AutoRegressive Integrated Moving Average.
Conv1D	1D Convolution.
DAI	Data-centric AI.
ELM	Extreme Learning Machine.
LSTM	Long Short-Term Memory.
MLP	Multi-Layer Perceptron.
RNA	Redes neurais Artificiais.
SARIMA	Seasonal AutoRegressive Integrated Moving Avegare.
SVR	Support Vector Regression.
SW	Sliding Window.

Capítulo I Introdução

O etanol é comercializado no Brasil na forma de etanol hidratado, utilizado diretamente como combustível veicular, ou como etanol anidro, que é adicionado atualmente em uma proporção de 27% na gasolina comum [ANP, 2022]. Além do Brasil, vários outros países utilizam o etanol como combustível veicular misturado à gasolina em percentuais que variam dependendo do país ou região. Entre esses países estão os EUA, países integrantes da União Europeia, China, Índia, Canadá, Tailândia e Argentina [Renewable Fuels Association, 2024]. Os principais motivadores para o uso do etanol misturado à gasolina incluem reduzir a dependência do petróleo ou a emissão de gases de efeito estufa, principalmente com percentuais mais elevados de etanol na mistura com a gasolina [Niven, 2005; Sadeghinezhad et al., 2014].

A importância do etanol como combustível para veículos leves no Brasil é bastante significativa. No ano de 2023, o volume total de etanol vendido no Brasil combustível foi de pouco mais de 28 milhões de m^3, o que correspondeu a quase 22% do volume total de combustíveis líquidos vendidos no país [ANP, 2024a,b]. A produção brasileira de etanol atualmente é majoritariamente baseada na utilização da cana de açúcar como matéria prima, embora a produção de etanol oriunda do milho tenha se intensificado em alguns estados brasileiros [Silva and Castañeda-Ayarza, 2021]. A produção brasileira de etanol se concentra em alguns poucos estados [ANP, 2024a], que devem atender às necessidades de consumo de todo o país.

Pelo fato da produção de etanol utilizar matérias primas de origem agrícola, possui um forte componente sazonal. Tal sazonalidade da produção representa um desafio logístico, uma vez que o consumo desse biocombustível não compartilha de tal sazonalidade. Além da sazonalidade, existem diversos outros fatores que podem influenciar a produção de etanol. Dentre eles podemos destacar (1) os fatores climáticos, que podem afetar a produtividade das safras de cana de açúcar e milho, (2) os preços de venda do etanol e do açúcar, que podem levar a decisão de priorizar um produto em detrimento do outro no caso da produção de etanol de cana de açúcar, (3) o preço do petróleo, que pode afetar a competitividade do etanol combustível perante a gasolina [Karp et al., 2021]. Mais recentemente a pandemia de COVID-19 também impactou significativamente a procura por combustíveis, incluindo o etanol, o que por consequência impacta a produção [Gao et al., 2022]. Apesar dessas influências, este estudo se baseia na suposição autorregressiva de que tais informações já estão incorporadas em termos autorregressivos [Gujarati, 2021], o que considera a modelagem

baseada apenas nas séries temporais mensais de produção de etanol, sem a utilização de variáveis exógenas.

Dada a sazonalidade inerente à produção, as variações na escala de produção entre os estados, a necessidade de armazenamento e transporte de etanol dos centros de produção até os consumidores finais e a influência de fatores climáticos e de mercado, a previsão da produção de etanol por unidade geográfica é comumente apoiada por modelos estatísticos e econométricos. Entretanto, identificou-se ainda que em menor escala, o uso de modelos específicos de inteligência artificial (Artificial Inteligence Model (AIM)) para a abordagem desse problema. Os Modelos de Inteligência Artificial podem melhorar o desempenho da previsão através do emprego de técnicas avançadas que permitem a captura de padrões complexos e relações não lineares dentro dos dados.

Notou-se, ainda, que pouca ênfase foi dada nos trabalhos à importância do pré-processamento de dados para auxílio ao desempenho de modelos preditivos. Tal abordagem centrada no tratamento dos dados de entrada, chamada de Inteligência Artificial Centrada em Dados (Data-centric AI (DAI)), pode representar uma contribuição importante para o melhor desempenho dos modelos preditivos [Jarrahi et al., 2023]. Outro detalhe observado nos trabalhos foi a ausência de abordagens comparativas das combinações de DAI e AIM mais eficientes para os problemas estudados.

Com base nas lacunas acima citadas, o problema de pesquisa pode ser articulado da seguinte forma: i) Como a utilização de AIM pode ser introduzida para apoiar previsões que superem modelos estatísticos bem estabelecidos? ii) As iniciativas DAI podem contribuir significativamente com tal objetivo? Este trabalho busca responder tais questões introduzindo o D-AI^2-M, uma metodologia DAI para melhorar a precisão da previsão e a eficiência dos AIM, melhorando assim a eficiência operacional logística.

Esta pesquisa tem como foco a utilização de AIM autorregressivos, aplicados às séries temporais mensais de produção de etanol nos principais estados produtores do Brasil. Buscou-se analisar os resultados obtidos sob duas perspectivas diferentes: i) análise de desempenhos *locais*, focada em identificar o melhor modelo para cada cenário (estado produtor e tipo de etanol produzido) individualmente, e ii) análise de desempenho *global*, que visa identificar o melhor modelo considerando o conjunto completo de todos os cenários estudados na avaliação experimental [Montero-Manso and Hyndman, 2021].

O restante desse trabalho está organizado em mais cinco capítulos. O capítulo II fornece uma revisão teórica dos principais conceitos e técnicas aplicáveis à previsão da produção de etanol. O Capítulo III investiga o corpo de trabalhos existentes na área. No Capítulo IV é apresentada a metodologia de pesquisa proposta. O Capítulo V descreva a avaliação experimental e discute os resultados obtidos. Por fim, no Capítulo VI são apresentadas as conclusões dessa pesquisa.

Capítulo II Referencial teórico

É apresentada nesse capítulo uma revisão teórica envolvendo os principais conceitos, métodos estatísticos, modelos de aprendizado e técnicas de preparação de dados aplicáveis na previsão da produção ou demanda de etanol combustível com base em modelos autorregressivos. Além disso o capítulo inclui uma análise de dados específicos relacionados com a situação da produção brasileira de etanol, apresentada a seguir.

II.1 Produção brasileira de etanol

A produção brasileira de etanol possui um componente sazonal significativo, com picos ocorrendo entre os meses de abril e novembro e os níveis mais baixos de produção ocorrendo entre os meses de dezembro e março. A variação mensal da produção brasileira de etanol é apresentada na Figura II.1, elaborada a partir de dados disponibilizados pela Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP) [ANP, 2024a].

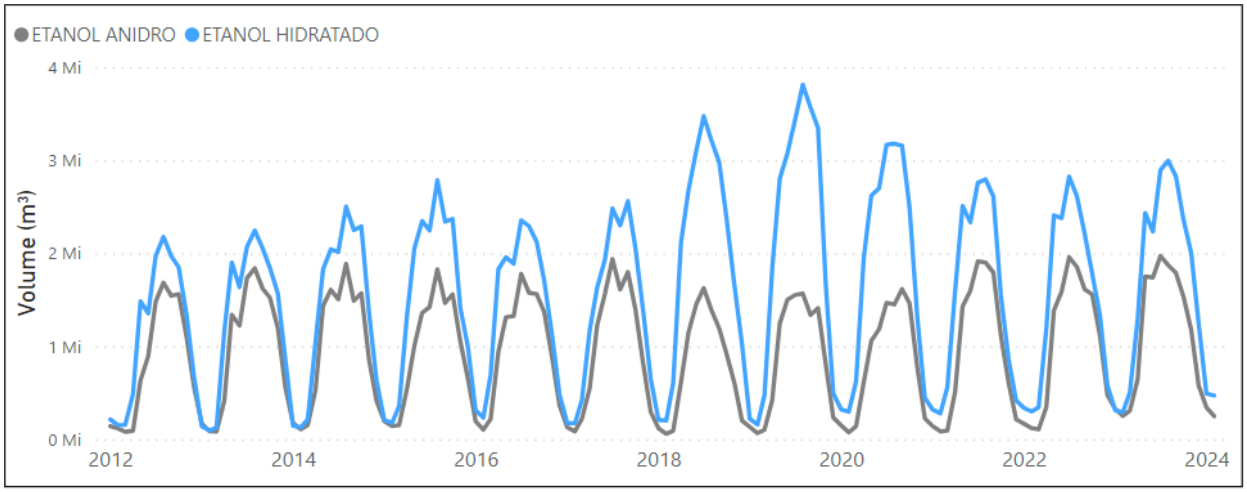


Figura II.1: Produção mensal brasileira de etanol anidro e hidratado (m^3), entre os anos de 2012 e 2023. Fonte: Adaptado de ANP [2024a]

Além do problema decorrente da sazonalidade, a produção de etanol combustível se concentra em alguns poucos estados brasileiros. Os estados de São Paulo (SP), Goiás (GO), Minas Gerais (MG), Mato Grosso (MT), Mato grosso do Sul (MS) e Paraná (PR) se destacam no volume de etanol produzido, representando juntos mais de 90% da produção brasileira conforme análise dos dados da ANP [2024a]. Tais estados devem atender à demanda contínua de consumo de todo o

país. A rede de distribuição e revenda de etanol é complexa e muitas vezes não se traduz em uma relação direta entre estado produtor e estado consumidor. Mesmo com isso em mente é apresentada na Tabela II.1 uma comparação ilustrativa dos volumes de etanol produzidos e consumidos por cada estado brasileiro entre 2019 e 2023. O volume de etanol anidro consumido foi estimado como 27% do volume da gasolina consumida no período. Analisando os números supracitados, podemos perceber o expressivo excesso de volume de etanol produzido em relação ao consumido pelos quatro maiores estados produtores (SP, GO, MT e MS).

Tabela II.1: Produção e consumo de etanol por estado brasileiro entre os anos de 2019 e 2023 (dados da ANP [ANP, 2024a,b])

Unidade da federação	**Produção (x1000 m^3)**	**Consumo (x1000 m^3)**
São Paulo (SP)	68.861	58.375
Goiás (GO)	26.350	9.319
Mato Grosso (MT)	19.025	5.332
Mato Grosso do Sul (MS)	15.706	1.783
Minas Gerais (MG)	15.696	17.256
Paraná (PR)	6.440	9.745
Alagoas (AL)	2.249	953
Paraíba (PB)	1.898	1.619
Pernambuco (PE)	1.720	3.114
Bahia (BA)	1.559	5.353
Tocantins (TO)	938	677
Maranhão (MA)	810	1.552
Rio Grande do Norte (RN)	560	1.182
Rio de Janeiro (RJ)	547	5.985
Espírito Santo (ES)	498	1.585
Sergipe (SE)	482	713
Pará (PA)	273	1.938
Piauí (PI)	221	1.137
Amazonas (AM)	37	1.486
Rondônia (RO)	5	668
Rio Grande do Sul (RS)	2	4.916
Acre (AC)	0	226
Amapá (AP)	0	248
Ceará (CE)	0	2.536
Brasília (DF)	0	2.110
Roraima (RR)	0	235
Santa Catarina (SC)	0	4.193

Outra característica importante da produção brasileira de etanol é a crescente utilização do milho como matéria prima. A série temporal de produção de etanol disponibilizada pela UNICAdata [UNICAdata, 2023] trás importantes informações sobre esse tema. Tal série temporal é de frequência anual, e segrega a produção de etanol pela matéria prima utilizada (cana de açúcar ou milho).

Segundo os dados da UNICAdata, os primeiros registros da produção brasileira de etanol a partir do milho ocorreram no ano de 2014. A produção de etanol de milho, embora ainda seja signi-

ficativamente inferior á produção de etanol de cana de açúcar, tem de se ampliado gradativamente conforme pode-se observar na Figura II.2. Esse detalhe é relevante pois a produção de etanol de milho não sofre a pressão econômica da variação do preço do açúcar, como ocorre com o etanol da cana-de-açúcar. Dessa forma, insere-se uma variável diferente na tomada de decisão referente ao volume de etanol a ser produzido.

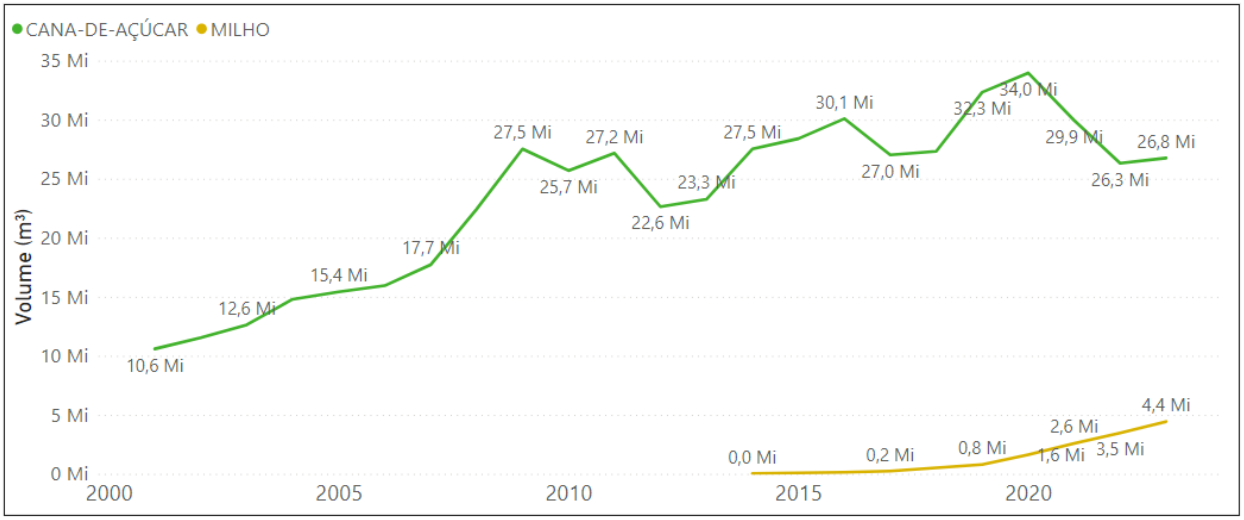


Figura II.2: Utilização de cana de açúcar x milho na produção brasileira de etanol. Fonte: Adaptado de UNICAdata [2023]

A Figura II.3 apresenta a variação, ao longo dos anos, do volume de etanol proveniente do milho e da cana de açúcar para os quatro estados que produzem etanol de milho. Destaca-se o estado do Mato Grosso, onde a produção de etanol de milho superou o etanol de cana de açúcar a partir de 2020 e o estado do Mato Grosso do Sul que no ano de 2023 registrou um primeiro, porém significativo volume de etanol de milho produzido.

Outro dado relevante para a interpretação do comportamento dos modelos é a variação histórica do percentual de etanol anidro adicionado na gasolina c brasileira. Tal variação é apresentadas na Tabela II.2, elaborada com informações do Ministério da Agricultura e Pecuária [da Agricultura e Pecuária, 2022]. Nota-se que a maior parte da série temporal disponibilizada pela ANP, que tem seu início em janeiro de 2012, contempla o período em que o percentual de etanol anidro na gasolina comum foi definido como 27%.

Tabela II.2: Mudanças nas regras para a mistura de etanol anidro na gasolina comum brasileira

Dispositivo legal	Data da edição	Vigência	Percentual de etanol na gasolina
Portaria MAPA nº 678	31/08/2011	01/10/2011	20%
Portaria MAPA nº 105	28/02/2013	01/05/2013	25%
Portaria MAPA nº 75	05/03/2015	16/03/2015	27%

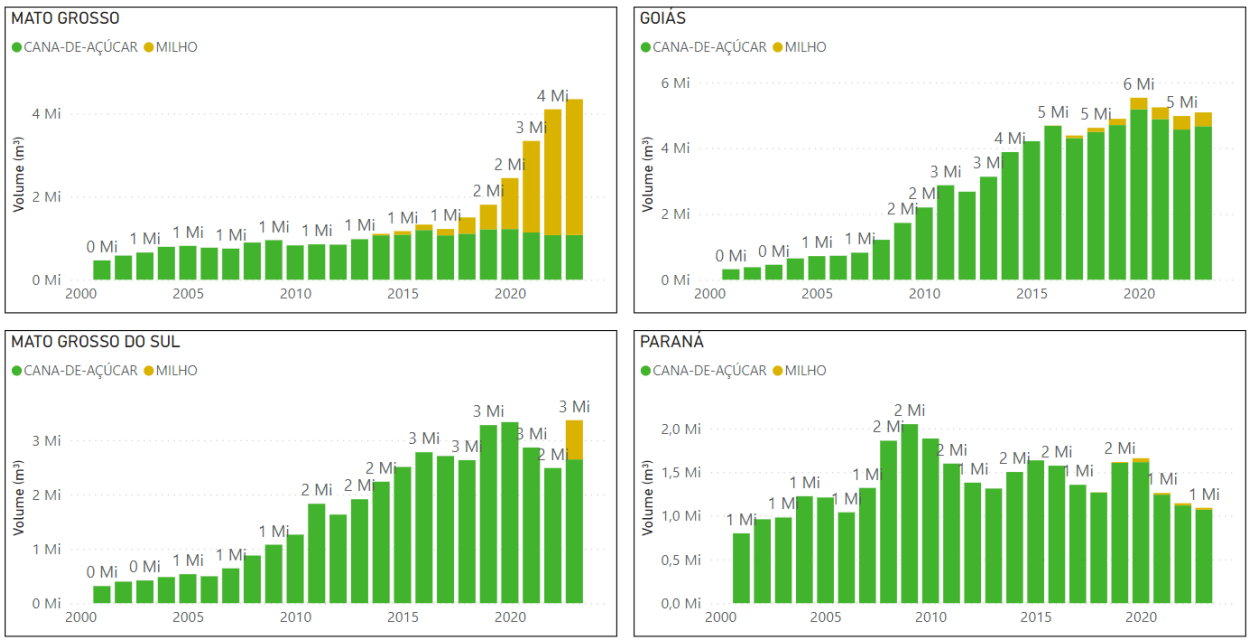


Figura II.3: Estados brasileiros que utilizam o milho como matéria prima para a produção de etanol - Comparação dos volumes produzidos. Fonte: Adaptado de UNICAdata [2023]

II.2 Séries temporais

Séries temporais são sequências de dados ordenadas e coletadas em intervalos de tempo regulares e conhecidos [Al-Fattah, 2020]. As séries temporais podem ter um ou mais dos seguintes componentes: i) Tendência: Comportamento observado quando a série temporal mostra um comportamento crescente ou decrescente ao longo do tempo, não necessariamente linear; ii) Sazonalidade: Um padrão de repetição notado dentro de uma certa periodicidade; iii) Ciclo: Padrões cíclicos que ocorrem como flutuações em torno da tendência devido a fatores econômicos, ambientais ou outros. Os ciclos não têm duração fixa e podem variar em duração, diferindo da sazonalidade[Hyndman and Athanasopoulos, 2018] e iv) Ruído ou erro: Parte aleatória da série temporal.

Uma série temporal pode ser decomposta segundo o modelo aditivo, apresentado na Equação II.1 ou segundo o modelo multiplicativo, apresentado na Equação II.2. Em ambas as equações y_t representa uma série temporal composta pela combinação das componentes de tendência β_t, sazonalidade s_t e erro e_t (parte aleatória não capturada). O modelo multiplicativo é apropriado quando a variação do componente sazonal ao redor do componente de tendência for proporcional à magnitude dos dados, enquanto o modelo aditivo é mais adequado quando tal relação não for proporcional [Barros et al., 2018].

$$y_t = \beta_t + s_t + e_t \tag{II.1}$$

$$y_t = \beta_t \times s_t \times e_t \tag{II.2}$$

Um exemplo de série temporal decomposta em seus diferentes componentes é apresentado na Figura II.4. A parte superior da figura apresenta a série temporal original (1) e em seguida os componentes de tendência (2), sazonalidade (3) e por fim o componente aleatório (4) ou ruído. O exemplo da Figura II.4 foi elaborado a partir de uma série temporal sintética TS, construída conforme apresentado na equação II.3. TS é o resultado da soma de componentes de tendência, sazonalidade e ruído aleatório respectivamente. Para simular o ruído aleatório e_t utilizou-se uma distribuição normal com média 0 e desvio padrão 5.

$$TS = 0.5t + 10\sin\left(\frac{2\pi t}{12}\right) + e_t \tag{II.3}$$

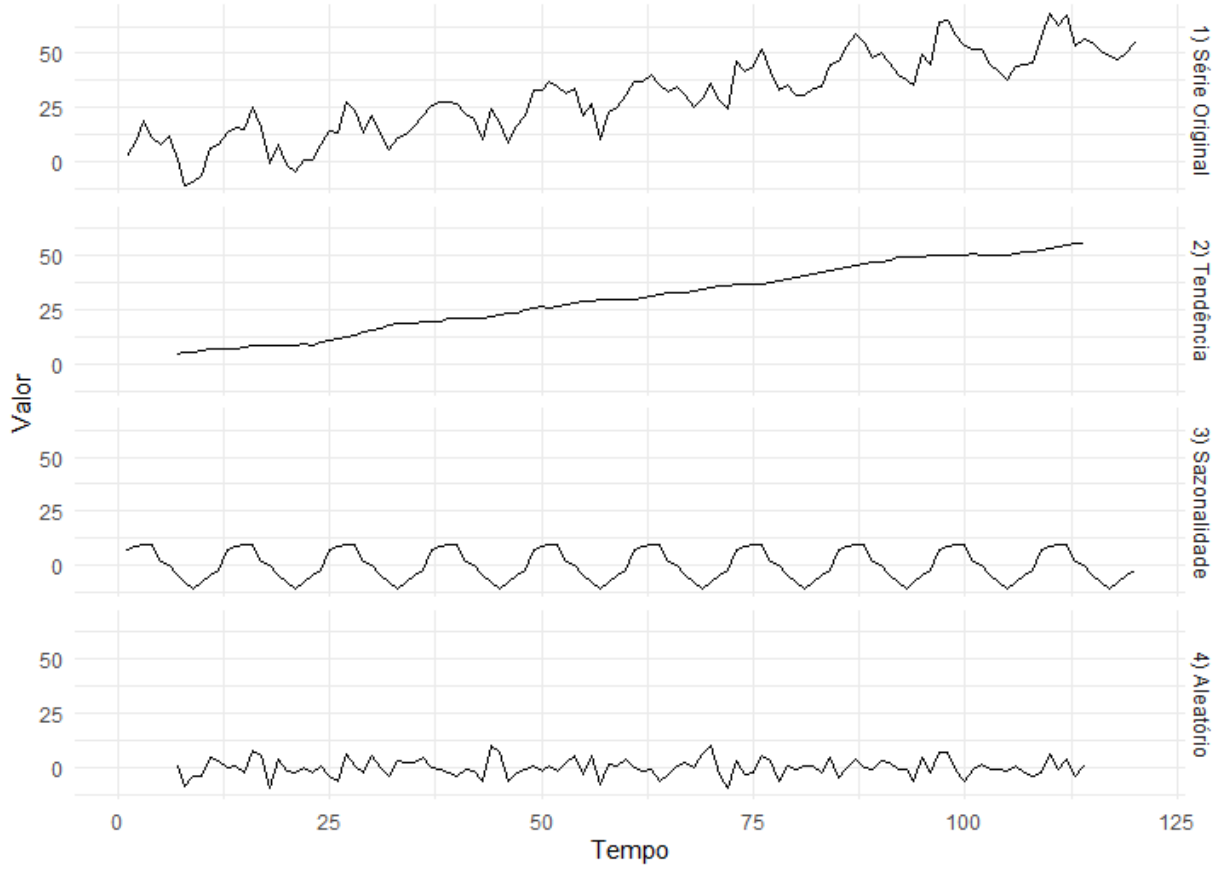


Figura II.4: Decomposição de uma série temporal em seus componentes

II.3 Modelos Auto-Regressivos

Modelos auto-regressivos utilizam os valores anteriores de uma determinada variável para prever seus valores futuros. A variável de interesse é regredida em seus próprios valores defasados (o que

explica o prefixo *auto* do termo auto-regressivo), de modo a se estimar seus valores numéricos posteriores. Essa classe de modelos busca identificar padrões de dependência temporal presentes nos dados para realizar as respectivas previsões [Dritsaki et al., 2021].

A aplicação de tal técnica considera que o conjuntos de dados a ser utilizado é uma série temporal. Esta pesquisa considera especificamente a utilização de modelos auto-regressivos na metodologia proposta. Nas próximas seções são apresentadas técnicas e algoritmos aplicáveis à modelos auto-regressivos.

II.4 Modelos ARIMA

Um dos modelos auto-regressivos mais frequentemente utilizados é o AutoRegressive Integrated Moving Average (ARIMA). Como o acrônimo ARIMA sugere, este modelo combina as seguintes técnicas: (AR) componente Auto-Regressivo, que assume que o valor atual de uma série temporal depende dos valores anteriores defasados da série temporal. O valor do termo p na representação do modelo ARIMA(p, d, q) indica quantos períodos anteriores estão sendo considerados; (I) componente de Integração, responsável pela diferenciação a ser realizada na série temporal para torná-la estacionária. O valor do termo d indica quantas vezes a série precisa ser diferenciada para alcançar a estacionariedade; (MA) componente de Média Móvel, que considera a média dos erros de previsões anteriores para modelar o comportamento da previsão atual (em vez de avaliar valores anteriores e o valor atual como faz o componente AR). O valor do termo q indica quantos erros passados são considerados no componente MA [Li et al., 2010; Nielsen, 2019]. Na notação tradicional, o modelo ARIMA pode ser representado como na Equação II.4 [Nielsen, 2019; Hyndman and Athanasopoulos, 2018].

$$y_t^d = \phi_0 + \sum_{i=1}^{p}(\phi_i \times y_{t-i}^d) - \sum_{i=1}^{q}(\theta_i \times e_{t-i}) + e_t \tag{II.4}$$

Na equação II.4, y_t^d é a série temporal diferenciada conforme a ordem d [Hyndman and Athanasopoulos, 2018]. Por exemplo, se d=1 (uma diferenciação) uma nova série temporal é criada subtraindo-se o valor atual y_t pelo valor imediatamente anterior y_{t-1} para todas as observações. Para d>1 o processo se repete conforme a ordem estabelecida por d. O termo ϕ_0 é uma constante também chamada de "média não nula". O termo $\sum_{i=1}^{p}(\phi_i \times y_{t-i}^d)$ é o componente autorregressivo (AR), de ordem p a ser determinada e cujos valores de ϕ podem ser estimados por técnicas de otimização que minimizem a soma dos erros quadrados das previsões. O termo $\sum_{i=1}^{q}(\theta_i \times e_{t-i})$ é o componente de média móvel (MA), de ordem q. Por fim, o termo e_t é a parte aleatória da série temporal.

A determinação dos parâmetros p, d e q do modelo ARIMA(p,d,q), quando ocorre de forma

manual, normalmente se dá através da análise de gráficos de autocorrelação (ACF) e autocorrelação parcial (PACF) [Nielsen, 2019]. Quedas abruptas no gráfico de ACF após um determinado atraso indicam a ordem apropriada dos termos MA (ordem q) e Quedas abruptas no gráfico de PACF após um determinado atraso indicam a ordem apropriada dos termos AR (ordem q). Entretanto, existem atualmente, incorporadas a diversas ferramentas de modelagem, recursos de ajuste automatizado de modelos ARIMA [Nielsen, 2019], sendo uma das mais relevantes a função *auto.arima*(), integrante do pacote *forecast* desenvolvido na linguagem R [Hyndman and Athanasopoulos, 2018].

Uma extensão do modelos ARIMA é o Seasonal AutoRegressive Integrated Moving Avegare (SARIMA), que foi desenvolvido para séries temporais que exibem padrões sazonais que devem ser considerados. O modelo SARIMA adiciona os seguintes componentes adicionais: (SAR) Componente Autorregressivo Sazonal; (SI) Componente de diferenciação Sazonal e (SMA) Componente de média móvel Sazonal. A representação do modelo SARIMA se dá na forma "SARIMA $(p,d,q)(P,D,Q)$" onde p, d e q representam os componentes não sazonais, similares aos do modelo ARIMA, enquanto P, D e Q representam, respectivamente, as ordens sazonais para os componentes auto-regressivos, de diferenciação e de média móvel do modelo [Figueira et al., 2010].

Pode-se destacar como uma das principais vantagens dos modelos ARIMA o baixo custo computacional necessário para a etapa de treinamento do modelo. Além disso, a quantidade limitada de parâmetros do modelo ARIMA facilita a sua análise e interpretação. Como ponto negativo pode-se destacar que modelos ARIMA não são tão eficientes ao lidar com dados complexos ou não lineares.

II.5 Data-centric AI (DAI)

DAI é um conceito que enfatiza a importância do tratamento e da melhoria dos dados de entrada para algoritmos AIM. Esta abordagem oferece a possibilidade de melhoria de desempenho, complementar ao refinamento dos algoritmos [Jarrahi et al., 2023].

Um importante pré-processamento DAI para suportar AIM é transformar uma série temporal em uma Sliding Window (SW), ou janela deslizante. Em vez de considerar o vetor de dados inteiro, essa abordagem o divide em segmentos de SW de comprimento n. Seja uma série temporal X uma sequência de n observações $< x_1, x_2, x_3, \ldots, x_n >$, onde x_1 representa o valor assumido pela série na observação mais antiga e x_n representa o valor da série na observação mais recente. Uma subsequência de tamanho p obtida de X que termina na posição i pode ser representada por $seq_{i,p}(X)$ que é equivalente à sequência de valores $< x_i - (p+1), x_i - (p+2), \ldots, x_i >$, onde $|seq_{i,p}| = p$ e $p \leq i \leq |X|$. SW é uma abordagem para explorar todas as possíveis sub-sequências de tamanho p de uma série temporal [Ogasawara et al., 2010].

Outra aplicação importante de preprocessamento DAI é a normalização de dados. Combinada com SW, a normalização pode melhorar significativamente os AIM. Existem muitos métodos de pré-

processamento disponíveis, como global min-max (GMM) [Ogasawara et al., 2009], diferenciação (DIFF) [Gujarati, 2021] e normalização adaptativa (AN) [Ogasawara et al., 2010].

GMM normaliza os valores de um dado vetor de variáveis baseado em seus valores máximos e mínimos atuais e os novos valores máximos e mínimos após o pré-processamento. Esse processo é ilustrado na equação II.5 onde x representa o vetor de dados original, x' o vetor normalizado, new_{max} representa o novo valor máximo, e new_{min} o novo valor mínimo do vetor normalizado [Ogasawara et al., 2010].

$$x' = (new_{max} - new_{min}) \times \frac{x - min_x}{max_x - min_x} + new_{min} \tag{II.5}$$

DIFF é um método usado para transformar uma série temporal não estacionária em uma série estacionária. Isso é feito calculando as diferenças entre valores consecutivos na série temporal. Esse processo pode ser repetido para obter a segunda diferença, a terceira diferença e assim por diante, até que a série se torne estacionária. Esta técnica é aplicada antes do treinamento do modelo para garantir que os dados de entrada sejam apropriados para o modelo em questão, melhorando assim a qualidade da previsão.

AN é um método de pré-processamento para séries temporais não estacionárias heteroscedásticas. Diferente das técnicas tradicionais de janela deslizante, AN transforma séries temporais em sequências de dados onde propriedades estatísticas globais derivadas de um conjunto de amostras são integradas para apoiar a normalização. Isso permite que AN represente efetivamente volatilidades variáveis dentro de suas janelas deslizantes. A metodologia de AN é estruturada em três estágios distintos: (i) transformar a série temporal não estacionária em uma sequência estacionária por meio da criação de janelas deslizantes disjuntas, (ii) a remoção de outliers, e (iii) a normalização propriamente dita. Essa abordagem oferece uma solução abrangente para lidar com as complexidades inerentes a dados de séries temporais não estacionárias heteroscedásticas [Ogasawara et al., 2010].

II.6 AIM

AIM abrange uma ampla gama de modelos baseados em dados. Dentre eles incluem-se o Multi-Layer Perceptron (MLP), Extreme Learning Machine (ELM), Long Short-Term Memory (LSTM), 1D Convolution (Conv1D), e Support Vector Regression (SVR). Algorítimos AIM são mais eficientes que modelos estatísticos ao lidar com dados mais complexos ou não lineares. Entretanto apresentam como desvantagens a necessidade de maior esforço computacional durante a etapa de treinamento, a maior dificuldade na interpretação de resultados e a possibilidade de sofrerem sobreajuste (overfitting) se não forem bem ajustados. As próximas subseções apresentarão as características principais

de tais modelos.

II.6.1 MLP e ELM

MLP são modelos computacionais classificados como Redes neurais Artificiais (RNA). Tais modelos, inspirados na estrutura dos cérebros, consistem em unidades de processamento interconectadas (chamadas *neurônios*), que trabalham juntas para analisar dados e fazer previsões. Cada neurônio aplica uma função, geralmente não linear, à soma ponderada dos sinais recebidos para determinar seu sinal de saída [Nasr et al., 2003]. Esta função é referida como função de ativação. As funções de ativação mais comuns são sigmóide, ReLu e softplus, representadas respectivamente nas equações II.6, II.7 e II.8 [Russell and Norvig, 2019].

$$Sigmoide(x) = \frac{1}{(1 + e^{-x})} \tag{II.6}$$

$$ReLU(x) = max(0, x) \tag{II.7}$$

$$softplus(x) = log(1 + e^{x}) \tag{II.8}$$

As funções de ativação supracitadas são representadas em um gráfico comparativo na Figura II.5. A função de ativação sigmoide leva a saída do neurônio para o valor um quando a sua respectiva entrada ponderada é positiva e grande o suficiente, e leva a saída para zero quando a entrada ponderada é negativa e grande o suficiente em módulo. A função ReLU faz com que a saída do neurônio seja zero para valores negativos da entrada ponderada e torna a saída linear (igual a entrada) para valores positivos da entrada ponderada. A função softplus é uma versão suavizada da função ReLU para valores próximos de zero.

MLP consistem em pelo menos três camadas de neurônios: uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Um exemplo simplificado de MLP é apresentado na Figura II.6, onde i) $X1$ e $X2$ são as entradas de dados, ii) as setas representam os envios de sinais com a aplicação dos respectivos pesos, iii) $a1$ e $a2$ representam os neurônios da camada oculta, que aplicam aos seus sinais de saída suas respectivas funções de ativação e iv) $y1$ é a saída da rede neural.

Durante o treinamento do modelo, o processo onde os sinais fluem da camada de entrada, passando pelas camadas intermediárias até chegar na camada de saída é conhecido como avanço (do inglês, *feedforward*). Ao final de cada ciclo *feedforward* o algoritmo calcula e realiza a retropropagação (*backpropagation*) do erro (diferença entre as saídas previstas e as saídas reais) e realiza ajustes nos pesos das conexões entre os neurônios. Esse processo se repete, normalmente suportado

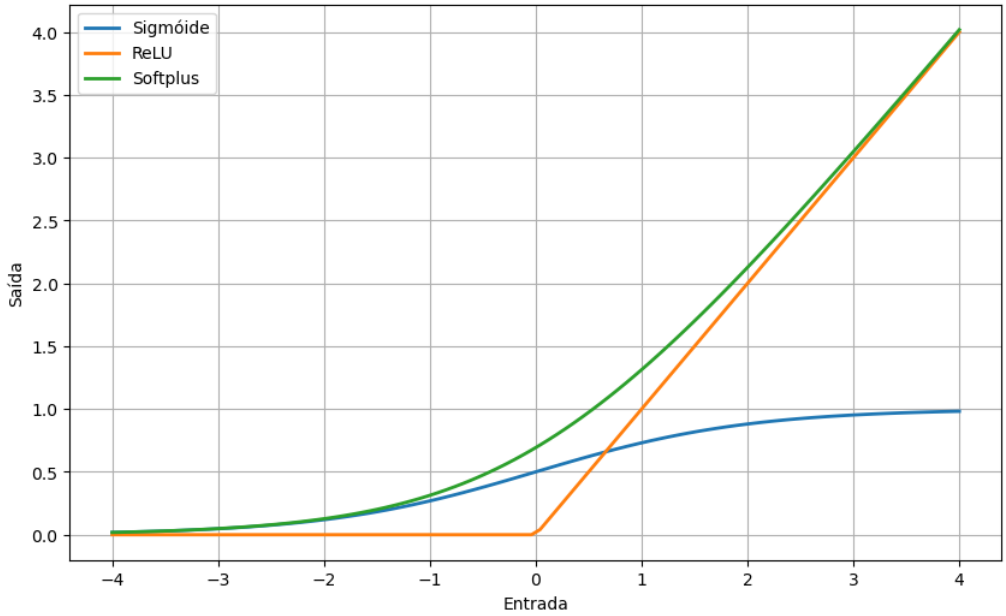


Figura II.5: Funções de ativação sigmoide, ReLU e softplus

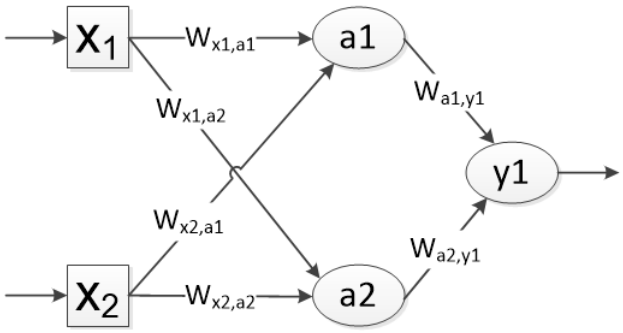


Figura II.6: Exemplo de MLP com duas entradas X_1 e X_2, uma camada oculta com dois neurônios a_1 e a_2 e um neurônio de saída y_1

por um algoritmo de otimização como o Gradiente Descendente, até que os valores ideais dos pesos sejam encontrados e o modelo seja considerado treinado [Anh Tran, 2021; Nasr et al., 2003].

O processo do Gradiente Descendente é apresentado no Algorítimo 1. Tal processo funciona ajustando iterativamente os pesos do modelo na direção da descida mais acentuada da função de perda de erro quadrático. A função de perda de erro quadrático mede a média dos quadrados dos erros, onde o erro é a diferença entre o valor previsto pelo modelo e o valor real observado. O parâmetro α no algorítimo representa a taxa de aprendizagem e pode possuir um valor constante ou pode decair conforme o processo de aprendizagem avança. A taxa de aprendizagem define o quão intensamente os valores dos pesos serão modificados a cada iteração. As iterações continuam até que um critério de parada seja satisfeito, tal como uma mudança pequena o suficiente na função perda ou o atingimento de um número máximo de iterações previamente estabelecido. No algorítimo, o objeto $Perda(w)$ representa a respectiva função de perda [Russell and Norvig, 2019].

Existe um tipo de rede neural artificial de camada oculta única, chamada de Extreme Learning

Algorithm 1 Gradiente Descendente

1: Inicialize os parâmetros w aleatoriamente
2: Defina a taxa de aprendizado α
3: **while** não convergir **do** ▷ Executa até que o critério de parada pré estabelecido seja satisfeito
4: Calcule o gradiente G, onde G = $\frac{\delta Perda(w)}{\delta wi}$
5: Atualize $w \leftarrow w - \alpha G$
6: **end while**

Machine (ELM), que possui a vantagem de ser treinada significativamente mais rapidamente que uma MLP. A característica principal de algoritmos ELM é que os pesos dos neurônios da camada oculta são iniciados aleatoriamente e permanecem fixos, e apenas os pesos da camada de saída são ajustados durante o treinamento do modelo, o que favorece a rapidez do treinamento [Gosso and Martinez-de Pison, 2012]. Apesar da simplicidade e rapidez do treinamento, os ELMs demonstram uma boa capacidade de generalização, sendo aplicáveis a uma ampla gama de tarefas, incluindo modelos autorregressivos.

II.6.2 Modelos de Aprendizado Profundo (LSTM e Conv1D)

Modelos de aprendizado profundo (do idioma inglês *deep learning*) também consistem em uma abordagem de aprendizado baseada no uso de redes neurais artificiais tais como MLP e ELM. Entretanto, ao contrário das redes neurais artificiais convencionais, os modelos de aprendizado profundo possuem várias camadas ocultas de neurônios, daí o termo *profundo*, que permitem que o algoritmo aprenda padrões mais complexos dos dados.

Nos trabalhos relacionados estudados, a estratégia de *deep learning* citada foi a Rede Neural Recursiva (em inglês, *Recursive Neural Network).* Redes neurais recursivas são modelos autoregressivos, utilizadas para resolver problemas relacionados com sequência ou tempo, e utilizam como dados as séries temporais. Embora redes neurais artificiais convencionais ou mesmo redes neurais profundas convencionais possam assumir que as entradas e as saídas são independentes, redes neurais recursivas dependem dos valores anteriores dos dados para a realização das predições [Manowska and Bluszcz, 2022].

Os trabalhos relacionados analisados citaram o uso de uma variação da arquitetura de redes neurais recursivas, chamada Long Short-Term Memory (LSTM), que inclui o uso de células de memória adicionadas à rede neural para armazenar informações relevantes por um longo período. Essas células de memória possuem um sistema de tomada de decisão que permite à rede *lembrar* ou *esquecer* informações, dependendo de sua relevância para a tarefa. Tais células de memória possuem três *portas*: i) porta de entrada: recebe os dados de entrada do modelo; ii) porta de esquecimento: decide o quanto da informação armazenada na célula de memória deve ser esquecida e iii) porta de saída: decide o quanto da informação armazenada na célula de memória deve ser transmitida para

a próxima célula [Marquez et al., 2021].

Outra técnica de aprendizado profundo com aplicações em séries temporais é a 1D Convolution (Conv1D), que incorpora a técnica de convolução no processamento de dados. A convolução aplica um filtro (ou kernel) para extrair características importantes de uma sequência de dados. Esse filtro desliza sobre a sequência de dados, o que pode capturar características específicas e padrões mais complexos de forma mais precisa. O Conv1D é capaz de não somente extrair as características mais relevantes dos dados sequenciais, mas também de diminuir a dimensão da sequência [Bhanja and Das, 2019].

II.6.3 Regressão de Suporte Vetorial (SVR)

O algorítimo Support Vector Regression (SVR) é uma extensão focada em regressão do algoritmo de Máquina de Vetores de Suporte (do inglês *Support Vector Machine* ou SVM). A principal vantagem do SVR em relação a outros métodos de regressão é a adoção do princípio da minimização do risco estrutural, que consiste minimizar um limite superior de erro de generalização ao invés do erro de treinamento, favorecendo assim a generalização do modelo [Li et al., 2022b].

O SVR busca encontrar um hiperplano que tenha o máximo número de pontos da saída y ao seu redor, dentro de uma margem pré determinada ϵ (chamada de tubo-ϵ), enquanto minimiza a magnitude dos pesos w (para reduzir a complexidade do modelo). Os pontos fora dessa margem representada na equação II.9 são considerados violações e são penalizados durante o treinamento do modelo [Yu et al., 2019]. A função do hiperplano SVR está representada na equação II.10 onde w é o vetor de pesos do hiperplano, x e o vetor de entradas independentes e b e o termo de viés (valor escalar de ajuste).

$$f(x) - \epsilon \leq y \leq f(x) + \epsilon \tag{II.9}$$

$$f(x) = w \cdot x + b \tag{II.10}$$

Para que a SVR encontre os hiperplanos de separação o algoritmo faz uso de funções chamadas de kernels. Tais funções transformam o espaço de entrada, normalmente em um espaço de dimensão mais alta. Isso é feito para que os dados que não são linearmente separáveis em seu espaço original possam se tornar separáveis no novo espaço. Os kernels mais comuns utilizados em SVR são: linear, polinomial, radial e sigmoide [Meyer et al., 2023].

A Figura II.7 apresenta um exemplo simples de modelo SVR ajustado. Para a elaboração de tal figura foi gerado um conjunto de dados sintéticos de forma senoidal com algum ruído propositalmente inserido. O gráfico resultante exibe os dados sintéticos originais representados como pontos de cor

laranja. A predição, por ser unidimensional, é representada pela linha azul (hiperplano obtido pelo modelo ajustado). O tubo-ϵ indicando a margem de erro previamente estabelecida ($\epsilon = 0,1$) é representado pela área azul ao redor da linha da predição.

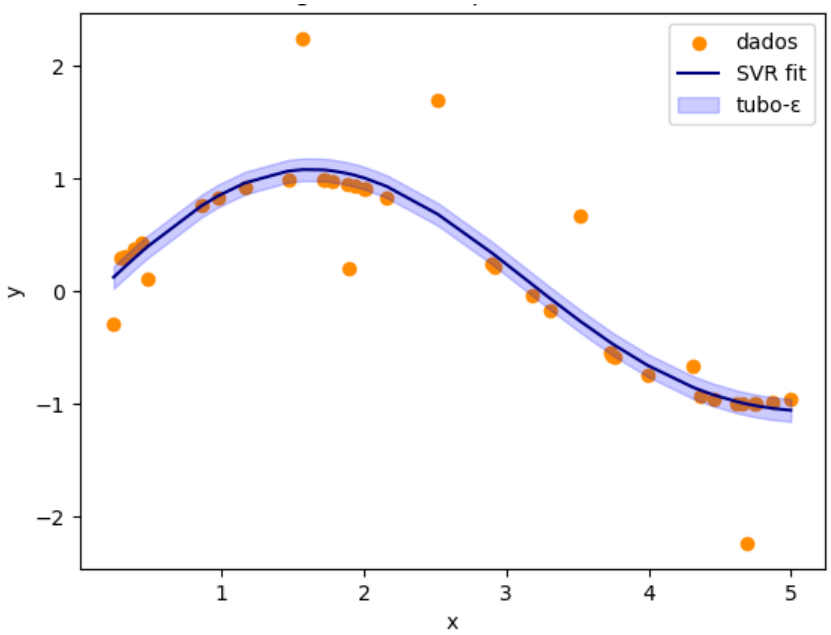


Figura II.7: Exemplo de Regressão de Suporte Vetorial

II.6.4 Modelos baseados em incerteza

Este item apresenta as técnicas de modelagem, identificadas nos trabalhos relacionados, que tiveram que lidar com problemas de incertezas e ambiguidade nos dados disponíveis. Duas abordagens de regressão foram identificadas: modelos Fuzzy e modelos Grey.

A regressão Fuzzy (difusa) é útil para estimar relações entre variáveis quando os dados disponíveis são imprecisos e as variáveis interagem de forma incerta. Ao contrário da regressão convencional, que é baseada na teoria da probabilidade, a regressão Fuzzy é baseada na teoria da possibilidade e na teoria do conjunto Fuzzy. Isso permite que as saídas do modelo sejam intervalos de valores com uma determinada amplitude. A amplitude do intervalo da saída reflete a influência do grau de incerteza ou imprecisão dos dados de entrada, o que possibilita uma representação mais abrangente e realista [Azadeh et al., 2014b, 2009].

Outra abordagem identificada nos trabalhos relacionados, para lidar com dados imprecisos, foi a regressão baseada no modelo Grey. O principal fundamento dessa técnica envolve um processo chamado de "operações de geração acumulada", que está relacionado com a etapa de preparação de dados para torna-los mais "cinzentos". Essa operação é realizada em duas etapas: a primeira é a geração acumulada original, que é usada para calcular a soma acumulativa dos valores da série original. A segunda etapa é a geração acumulada inversa, que é usada para gerar uma série de

dados diferenciados a partir dos dados acumulados. Esses dados diferenciados são então usados para construir equações diferenciais que modelam o comportamento do sistema e preveem seu comportamento futuro. As operações de geração acumulada ajudariam a lidar com a incerteza dos dados de entrada [Chen et al., 2021; Yang et al., 2016; Li et al., 2022a].

II.7 Otimização de hiperparâmetros

Hiperparâmetros são parâmetros definidos antes do início do processo de treinamento dos modelos, e que embora influenciem o desempenho do modelo, não podem ser aprendidos a partir dos dados. As opções de hiperparâmetros disponíveis dependem do tipo de modelo que será utilizado. Exemplos de hiperparâmetros incluem a taxa de aprendizagem e o número de camadas ocultas em uma rede neural, dentre outros [Mardiana et al., 2020]. Os próximos parágrafos apresentam uma introdução teórica sobre as principais técnicas de otimização de hiperparâmetros identificadas nos trabalhos relacionados.

(i) Pesquisa aleatória (*Random Search*): Define-se um intervalo de valores de hiperparâmetros para pesquisar. Para cada hiper-parâmetro, um valor aleatório é amostrado do intervalo definido. O modelo é então treinado e tem seu desempenho registrado. O processo é repetido por um número especificado de iterações ou até que o conjunto ideal de hiperparâmetros seja identificado [Gkerekos et al., 2019; Mardiana et al., 2020].

(ii) Pesquisa em grade (*Grid Search*): Consiste em testar todas as combinações possíveis de valores para os hiperparâmetros do modelo, dentro de um espaço de busca pré-definido, visando escolher a combinação que apresentar o melhor desempenho. Suponha um modelo de aprendizado de máquina que possua dois hiperparâmetros: a) Tamanho do vetor de entrada, com as seguintes opções pré definidas: 3, 4 ou 5 e b) Função de ativação, com as seguintes opções pré definidas: sigmóide, ReLU e Softplus. Em um processo de *Grid Search* todos as nove combinações possíveis serão testadas. Embora *Grid Search* seja uma técnica de implantação simples, pode ser computacionalmente custosa se a combinação de hiperparâmetros a ser testada for muito ampla [Marquez et al., 2021; Mardiana et al., 2020].

(iii) Otimização Bayesiana: Cria um modelo probabilístico da função objetivo (desempenho do modelo que se quer otimizar) que mapeia os hiperparâmetros para o desempenho do modelo. O modelo é atualizado à medida que novas avaliações da função objetivo são feitas e é utilizado para selecionar o próximo conjunto de hiperparâmetros a ser avaliado. Esse processo se repete até que um critério de parada seja satisfeito (como por exemplo um número máximo de iterações ou nenhuma melhora significativa de desempenho após várias iterações). Esse processo de seleção equilibra a exploração de novos hiperparâmetros com os melhores valores encontrados até o momento [Mardiana et al., 2020].

Capítulo III Trabalhos relacionados

Foi realizada uma busca sistemática de trabalhos relacionados seguindo às orientações da metodologia PRISMA 2020 [Page et al., 2021]. Tal revisão teve como objetivo identificar quais os dados, modelos e metodologias de predição estão sendo aplicados com maior sucesso para a resolução de dois problemas: i) a predição da produção de etanol combustível por região geográfica e ii) a predição de consumo de gasolina ou etanol combustível por região geográfica. A predição de consumo de gasolina ou de etanol foi analisada com o objetivo de identificar as sinergias entre a predição da demanda por tais combustíveis e a predição da produção de etanol anidro e hidratado para o atendimento das regiões demandantes.

As etapas realizadas durante a revisão de trabalhos relacionados estão representadas na Figura III.1. A primeira etapa, ocorrida no dia 8 de abril de 2023, envolveu a realização de consultas de referências nas bases de dados Scopus e Web of Science. As consultas ocorreram através de expressões de busca, apresentadas na tabela III.1, que tiveram o objetivo de selecionar trabalhos que atendessem a um dos seguintes critérios: i) Abordar o uso de modelos matemáticos ou técnicas de aprendizado de máquina na predição da produção ou consumo de etanol ou ii) Abordar o uso de modelos matemáticos ou técnicas de aprendizado de máquina na predição de consumo de gasolina. A inserção das palavras *gas* e *petrol* nas expressões de busca teve o objetivo de identificar de forma mais completa artigos que abordassem soluções de predição de consumo para gasolina, uma vez que tais palavras, dependendo principalmente do país onde são empregadas, podem se referir à gasolina.

Tabela III.1: Consultas realizadas nas bases Scopus e Web of Science

Base	Expressão de busca
Scopus	(TITLE-ABS-KEY (("ethanol consumption"OR "ethanol demand"OR "ethanol production"OR "gasoline consumption"OR "gasoline demand"OR "gas consumption"OR "gas demand"OR "petrol consumption"OR "petrol demand"AND "fuel")) AND TITLE-ABS-KEY (("model"OR "machine learning"OR "data mining")) AND TITLE-ABS-KEY (("forecast*"OR "predict*")) AND LANGUAGE (english))
Web of Science	ALL=(("ethanol consumption"OR "ethanol demand"OR "ethanol production"OR "gasoline consumption"OR "gasoline demand"OR "gas consumption"OR "gas demand"OR "petrol consumption"OR "petrol demand") AND "fuel"AND ("model"OR "machine learning"OR "data mining") AND("forecast*"OR "predict*"))

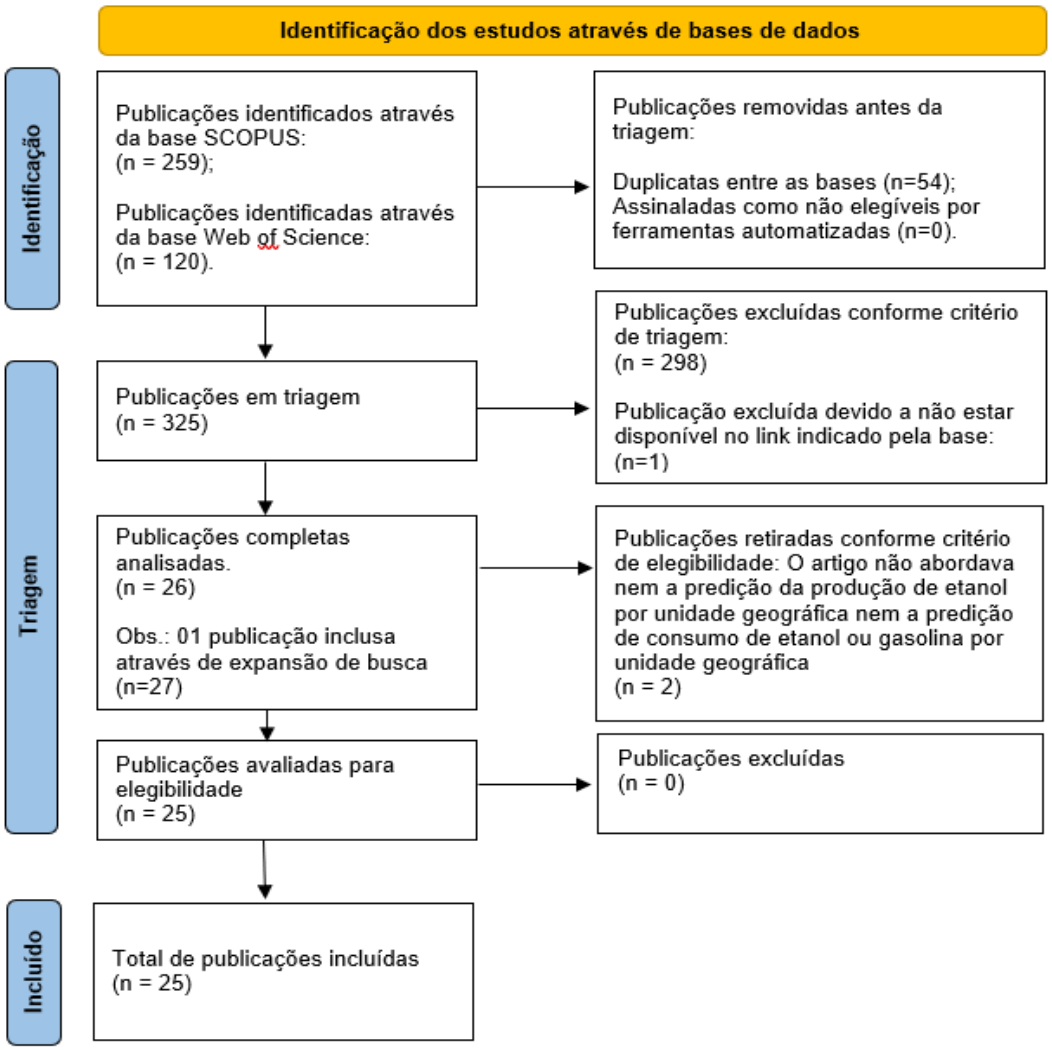


Figura III.1: Aplicação do fluxograma PRISMA para a seleção das referências a partir das bases [Page et al., 2021]

A base Scopus retornou 259 trabalhos enquanto a base Web of Science retornou 120 publicações. Foram identificados 54 trabalhos que constaram simultaneamente em ambas as bases. Desta forma foram colocados em triagem um total de 325 publicações. O critério de triagem adotado foi selecionar artigos que abordassem ao menos um dos cenários a seguir: i) O uso de modelos matemáticos e/ou técnicas de aprendizado de máquina para a predição de consumo de gasolina ou etanol, por região geográfica (estado/província, região ou país) ou ii) O uso de modelos matemáticos e/ou técnicas de aprendizado de máquina para a predição da produção de etanol por região geográfica (estado/província, região ou país). A triagem foi realizada a partir da leitura dos títulos e resumos dos trabalhos identificados. Nessa etapa 298 publicações foram excluídas por não atenderem aos critérios anteriormente citados.

Na etapa de leitura completa dos artigos foi excluída mais uma publicação, por não abordar nem a predição de consumo de gasolina ou etanol nem a predição de produção de etanol por região geográfica. Além disso, uma das publicações colocadas em triagem não pode avançar para a fase

de leitura de publicações completas devido a um problema no link para a mesma, indicado na base. Visando ampliar a quantidade de publicações abordando a produção de etanol (ou outros biocombustíveis) por unidade geográfica, foi conduzida uma expansão de busca a partir das referências inicialmente selecionadas. Tal expansão foi realizada através da avaliação dos trabalhos citados ou derivados a partir do conjunto inicial de referências obtidas. Nessa etapa foi identificado o artigo de Yu et al. [2022], que aborda a predição da produção de biocombustíveis nos EUA, e que foi incluso no estudo que passou a contemplar vinte e cinco trabalhos relacionados.

A Tabela III.2 apresenta o detalhamento do tipo de predição abordado em cada publicação selecionada, o tipo de método preditivo utilizado, o país onde tal predição foi aplicada e os *estratos Qualis* dos respectivos periódicos (Qualis calculado com dados da base Scopus, do ano de 2021). Na coluna *Predição* da Tabela III.2 a indicação *(C)* mostra que a predição é para consumo do combustível enquanto a indicação *(P)* mostra que a predição é de produção de etanol ou outros biocombustíveis. Analisando tal tabela podemos constatar que a maioria dos trabalhos relacionados aplica modelos estatísticos na resolução das predições abordadas (doze das vinte e cinco publicações). Entretanto, o uso de AIM é abordado em nove trabalhos, sendo sete abordando o uso de AIM isoladamente e dois abordando AIM em métodos combinados (mais de um algorítimo).

A análise das estratégias de predição, citadas pelo autores estudados é apresentada nos próximos itens deste capítulo da seguinte forma: o Item III.1 apresentará os trabalhos que envolveram a predição da produção de etanol ou outros biocombustíveis por região geográfica e o Item III.2 apresentará as publicações cujo foco foi a predição de consumo de gasolina ou de etanol (como combustível automotivo principal ou como aditivo para a gasolina) por região geográfica.

III.1 Predição da produção de etanol ou outros biocombustíveis

Fink and Medved [2011] propuseram o uso de modelos matemáticos para estimar como a temperatura do ar pode afetar a produção de culturas que servem como matéria-prima para a produção de etanol e biodiesel. Um aumento na temperatura impacta significativamente negativamente a produção de matérias-primas para o etanol e uma parcela substancial das matérias-primas utilizadas na produção de biodiesel. Eles propuseram uma equação para calcular a quantidade de combustível obtida a partir da cana-de-açúcar. O mesmo estudo concluiu que os efeitos da precipitação não teriam impacto significativo na produção de biocombustíveis.

MY and GS [2013] propuseram a utilização do modelo Hubbert para predizer a curva de produção de etanol. O modelo Hubbert é um método matemático que é comumente utilizado para prever a produção de petróleo de uma determinada região e segue um padrão de crescimento e declínio de produção em forma de curva assimétrica. Os autores propõe estender o uso de tal método para predições relacionadas com recursos renováveis, e desenvolvem um estudo focado na produção de

etanol nos EUA.

Badamchizadeh et al. [2021] propuseram um modelo autorregressivo baseado em uma rede neural artificial com 2 camadas (6 e 3 neurônios) para prever uma possível necessidade de produção anual de etanol no Irã, considerando 3 diferentes cenários de mistura de etanol na gasolina (uma vez que tal mistura não acontece nesse país). Os autores fizeram uma predição indireta de demanda por etanol em função da demanda por gasolina, e em seguida apresentaram alguma opções, potencialmente viáveis, de fontes de matérias primas para a produção de etanol a partir de resíduos da produção agrícola do Irã.

Yu et al. [2022] propõe uma combinação de métodos para a predição mensal da produção de biocombustíveis nos EUA. A metodologia proposta envolve quatro etapas: i) aplicação de EMD (*Empirical Mode Decomposition*), um método de decomposição dos dados originais em componentes mais simples; ii) utilização da técnica de LSTM para prever a componente de alta frequência da série temporal; iii) utilização da técnica ELM para prever a componente de baixa frequência da série temporal; iv) Integração dos resultados individuais de previsão em uma saída agregada por meio de adição simples. os autores compararam a abordagem combinada proposta com diferentes modelos de previsão, obtendo melhores resultados nas métricas RMSE e MAPE.

III.2 Predição de consumo de gasolina e/ou etanol combustível

Melikoglu [2014] propôs em seu artigo a utilização de modelos semi-empíricos para prever, dentre outras, a demanda por gasolina e etanol combustível na Turquia. A série temporal de consumo anual de gasolina na Turquia foi utilizada para realizar ambas as predições (demanda por gasolina e por etanol a ser adicionado na gasolina). Para o caso da predição de consumo de etanol, foi considerado o percentual previsto de mistura desse biocombustível na gasolina daquele país.

Wong et al. [1977] estimaram, com o uso de modelos econométricos, o consumo de gasolina no Estado de Minnesota, nos EUA, com base em dados de renda, preço da gasolina, divisão do mercado em veículos pequenos e grandes, eficiência energética e quilometragem percorrida. Os autores concluíram que mudanças no nível de renda são importantes para prever a demanda por energia automotiva no estado. Badr et al. [2008] investigaram o consumo de gasolina no Líbano, também fazendo uso de modelos econométricos, considerando o preço desse combustível e os registros de automóveis naquele país. Os autores confirmaram a significância estatística do preço da gasolina no consumo da gasolina, mas não confirmaram a significância estatística da série temporal de registro de automóveis.

Jeon [2019] fez uso de "Modelos de Circulação Geral" (modelos matemáticos complexos utilizados para simular o comportamento do sistema climático da Terra) para estimar o impacto das mudanças climáticas no consumo de combustíveis automotivos nos EUA. O autor concluiu que dias quentes

aumentam o consumo de gasolina, entretanto não há efeito estatisticamente significativo em dias frios, ao contrário do que ocorre no consumo de energia residencial.

Jaber et al. [2012] e Al-Ghandoor et al. [2013] realizaram predições baseadas em regressões lineares, onde ambos os trabalhos utilizam o número estimado de veículos leves, nível de renda e preço unitário da gasolina em suas predições. O primeiro trabalho constatou que sem a introdução de automóveis movidos à diesel o consumo de gasolina aumentaria 88,8% em um intervalo de 10 anos na Jordânia e devido a este fato encoraja tal mudança. O segundo artigo estima que o consumo de gasolina na Jordânia aumentará em uma taxa de 1,81% ao ano.

O trabalho de Figueira et al. [2010] utilizou o modelo SARIMA para prever o consumo de etanol hidratado com base na série histórica anual de consumo desse biocombustível no Brasil. Para prever o consumo de etanol anidro os autores utilizaram o conceito de função de transferência, que é uma extensão do modelo ARIMA que incorpora variáveis exógenas. Para esse caso os autores utilizaram a série histórica anual de consumo de gasolina como variável a ser modelada e a série história do PIB per capta anual brasileiro como uma variável exógena (explicativa). Com base em tal predição foi estimado o consumo de etanol anidro. Nesse processo, os autores incluíram a avaliação da correlação cruzada entre a variável exógena (PIB) e a variável dependente (consumo de gasolina) na seleção do melhor modelo. O artigo de Dey et al. [2023] também avaliou que o modelo ARIMA foi o mais preciso para prever a demanda anual por gasolina na Índia, e tal predição foi utilizada para estimar a demanda por etanol naquele país. Esse estudo compara o modelo ARIMA com os seguintes modelos matemáticos de regressão: regressão linear, regressão exponencial, regressão de potência, regressão quadrática, regressão inversa, regressão logarítmica e regressão S).

Cervero [1985], Waheed Bhutto et al. [2017] e Li et al. [2010] apontam os modelos ARIMA como os mais adequados nas suas predições de consumo de gasolina, sendo que o primeiro trabalho abordou o consumo mensal de gasolina nos Estados Unidos, o segundo o consumo anual de gasolina no Paquistão e o terceiro o consumo trimestral de petróleo para o segmento de transporte automotivo na Austrália. O trabalho de Cervero [1985] foi a publicação mais antiga identificada, dentre os trabalhos relacionadas a esta pesquisa, a mencionar a utilização de um modelo ARIMA. O artigo de Waheed Bhutto et al. [2017] estimou o consumo de gasolina no Paquistão com o objetivo de apoiar a formulação de políticas que favoreçam o acréscimo de etanol na gasolina desse país. A publicação de Li et al. [2010], embora indique o uso de modelos ARIMA para a predição supracitada, recomenda o uso de modelos mais simples se o objetivo for compreender melhor a influência sofrida pela variável dependente. Mardiana et al. [2020] realizaram uma comparação entre diversos modelos preditivos e concluiu que modelo de Holt-Winters obteve o melhor desempenho na predição de consumo anual de gasolina na Indonésia.

Kialashaki and Reisel [2014], Al-Fattah [2020] Mubarak et al. [2022] e Nasr et al. [2003] de-

senvolveram suas predições de consumo de gasolina baseadas em redes neurais artificiais, onde o primeiro trabalho utilizou PIB, população, preço do petróleo e número de veículos como variáveis independentes. O segundo trabalho utilizou como dados de entrada o PIB, preços da gasolina, preços do diesel, quantidade de veículos leves, quantidade de veículos pesados e dados de eficiência energética. O trabalho de Al-Fattah [2020] mencionou a realização de preprocessamento de dados (normalização) e também explicitou a utilização de algorítimo genético na estratégia de otimização de hiperparâmetros. O artigo de Mubarak et al. [2022] citou a utilização de re-amostragem e escalonamento como técnicas de pré-processamento. Para re-amostragem, os autores usam abordagens de interpolação de função de base radial (linear, cúbica e multi-quadrática) para expandir os dados brutos em um novo conjunto de dados. Isso foi feito para aumentar a quantidade de amostras de dados de treinamento para o algoritmo de aprendizado.

Marquez et al. [2021] apresentou em seu trabalho duas abordagens de utilização da técnica de *deep learning* LSTM para a predição de consumo mensal de etanol no Brasil: modelagem com múltiplas variáveis e modelagem univariada. Na modelagem univariada a série histórica de consumo mensal de etanol hidratado foi utilizada. Já na modelagem multivariada as seguintes séries temporais foram utilizadas: i) consumo mensal de etanol; ii) consumo mensal de gasolina; iii) frota mensal de veículos flex-fuel; iv) frota mensal de veículos à gasolina; v) preço médio mensal ponderado do etanol para o consumidor final e vi) preço médio mensal ponderado da gasolina para o consumidor final. Embora os autores tenham obtido os melhores resultados com o modelo LSTM multivariável, a previsão proposta foi realizada com o modelo LSTM univariado. Isso se deu, segundo os autores, devido à dificuldade de prever o comportamento temporal das variáveis exógenas (independentes).

O estudo apresentado por Rodrigues et al. [2022] avaliou o desempenho de vários modelos visando identificar qual o melhor método para prever o consumo mensal de combustíveis leves no Brasil (gasolina, etanol hidratado e gás natural veicular). Os modelos avaliados foram i) SARIMA; ii) ARIMA com variáveis binárias para sazonalidade; iii) Modelo vetorial de correção de erros (VECM), iv) Modelos de redes neurais artificiais univariados; v) Modelos de redes neurais artificiais multivariados; vi) SVR; vii) Sistema de inferência neural fuzzy híbrida (HyFIS), viii) Random Forest e, ix) Combinação de modelos (*ensemble learning*). Para a combinação de modelos os autores propuseram duas metodologias, todas levando em conta os resultados de todos os métodos de predição supracitados: A primeira estima os pesos para cada métodos por meio de mínimos quadrados restritos, impondo que a soma dos pesos seja igual a um. No segundo método, uma média simples das previsões dos diferentes modelos é tomada como a projeção final dos dados. Os resultados apresentados apontam que os melhores desempenhos foram alcançados pela combinação de modelos. Em relação ao desempenho individualizado dos modelos, os autores identificaram que o SVR, seguido de RF e ARIMA com variáveis binárias para sazonalidade apesentaram os melhores

desempenhos para a previsão no horizonte de 12 meses.

Yu et al. [2021] realizaram a predição anual de consumo de gasolina na China fazendo uso de uma técnica definida como "decomposição"na etapa de pré-processamento. Tal metodologia consistem em decompor a série temporal e realizar a predição de cada componente separadamente. Os autores fizeram as predições individuais baseadas em modelos SVR. O resultado final da predição consistiu na combinação dos modelos SVR através da soma simples de suas predições individuais.

Azadeh et al. [2014a] previram a demanda por gasolina nos Estados Unidos, Canadá, Japão, Irã e Kuwait utilizando modelos de regressão Fuzzy e os dados do PIB, população, número de veículos e preço da gasolina como variáveis independentes. Segundo os autores os modelos Fuzzy foram comparados com modelos de regressão convencionais através de analise de variância e erro percentual médio absoluto (MAPE). Os autores mencionaram ambiguidade nos dados obtidos relacionados com o consumo de gasolina, o que justificaria o melhor desempenho da regressão Fuzzy.

Chen et al. [2021] utilizaram um modelo baseado em uma melhoria em relação ao modelo Grey tradicional, chamado FGM (*Fractional Grey Model*), para prever o consumo anual de gasolina na China. O FGM se propõe a selecionar de maneira mais adequada o número de ordem de acumulação dos dados de entrada, o que, segundo os autores contribuiria com a obtenção de melhores resultados nas predições.

III.3 Discussão sobre os trabalhos relacionados

O conjunto mais pertinente de trabalhos relacionados ao tema central deste artigo está localizado na Subseção III.1, que se concentra na previsão da produção de etanol ou de outros biocombustíveis. Em contraste, a Subseção III.2 concentra-se na previsão do consumo de gasolina ou etanol combustível, empregando predominantemente modelos preditivos multivariados, com alguns estudos também integrando modelos autorregressivos.

No conjunto dos vinte e cinco trabalhos relacionados identificou-se que: i) a utilização de modelos estatísticos foi abordada em doze publicações (ARIMA, com cinco publicações, Regressões Lineares, com seis publicações, Holt-Winters, com uma publicação); ii) AIM foi utilizado em sete publicações (RNA convencional, com cinco publicações, LSTM e SVR, com uma publicação cada); iii) Combinação de métodos foi abordada em duas publicações e iv) outros métodos foram abordados em quatro publicações distintas. Nota-se, então, que ainda há espaço para o estudo da aplicação de modelos de inteligencia artificial (AIM) na predição da produção de etanol combustível por unidade geográfica.

Notou-se ainda uma significativa falta de ênfase relacionada às técnicas DAI empregadas para melhorar o desempenho dos modelos desenvolvidos nos trabalhos relacionados. Essa falta de ênfase foi identificada em todas as categorias de trabalhos selecionados, tanto no subconjunto de publi-

cações focadas na predição de consumo de gasolina ou etanol quanto no subconjunto de trabalhos focados na predição de produção de etanol ou outros biocombustíveis.

Embora os trabalhos relacionados apontem uma tendência de ampliação na utilização de AIM em problemas que envolvem a predição da produção e consumo de etanol por unidades geográficas, não foram identificados estudos comparativos amplos que explorem a seleção dos algoritmos mais adequados aos cenários estudados. Tampouco foram identificados estudos que envolvessem, de forma abrangente, a comparação das técnicas de DAI que mais contribuíram para o melhor desempenho dos modelos preditivos. A metodologia proposta para essa pesquisa, apresentada no próximo capítulo, visa preencher essas lacunas.

Tabela III.2: Trabalhos relacionados com a pesquisa proposta

Ano	Referência	Predição	Método preditivo	País	Qualis
2023	[Dey et al., 2023]	(C) gasolina, etanol adic.	ARIMA	Índia	A3
2022	[Mubarak et al., 2022]	(C) gasolina	RNA	Iraque	A1
2022	[Rodrigues et al., 2022]	(C) gasolina, etanol, GNV	Comb. métodos	Brasil	A4
2022	[Yu et al., 2022]	(P) etanol	Comb. métodos	EUA	A2
2021	[Badamchizadeh et al., 2021]	(P) etanol	RNA	Irâ	A4
2021	[Chen et al., 2021]	(C) gasolina	Incerteza	China	A2
2021	[Marquez et al., 2021]	(C) etanol	LSTM	Brasil	A4
2021	[Yu et al., 2021]	(C) gasolina	SVR	China	A2
2020	[Al-Fattah, 2020]	(C) gasolina	RNA	Arábia Saudita	A1
2020	[Mardiana et al., 2020]	(C) gasolina	(Outros) Holt-Winters	Indonésia	A1
2019	[Jeon, 2019]	(C) gasolina	(Outros) Circulação Geral	EUA	A2
2017	[Waheed Bhutto et al., 2017]	(C) gasolina	ARIMA	Paquistão	A3
2014	[Azadeh et al., 2014b]	(C) gasolina	Incerteza	EUA, Canadá, Japão, Irã e Kuwait	B2
2014	[Kialashaki and Reisel, 2014]	(C) gasolina	RNA	EUA	-
2014	[Melikoglu, 2014]	(C) gasolina, etanol adic.	Regressivo (semi-empíricos)	Turquia	A1
2013	[Al-Ghandoor et al., 2013]	(C) gasolina	Regressivo (lineares)	Jordânia	A1
2013	[MY and GS, 2013]	(P) etanol	(Outros) Hubbert	EUA	-
2012	[Jaber et al., 2012]	(C) gasolina	Regressivo (lineares)	Jordânia	B2
2011	[Fink and Medved, 2011]	(P) etanol	Regressivo (lineares)	Turquia	-
2010	[Figueira et al., 2010]	(C) etanol	ARIMA	Brasil	A3
2010	[Li et al., 2010]	(C) gasolina	ARIMA	Austrália	A1
2008	[Badr et al., 2008]	(C) gasolina	Regressivo (econométricos)	Líbano	A2
2003	[Nasr et al., 2003]	(C) gasolina	RNA	Líbano	A1
1985	[Cervero, 1985]	(C) gasolina	ARIMA	EUA	-
1977	[Wong et al., 1977]	(C) gasolina	Regressivo (econométricos)	EUA	A4

Capítulo IV Metodologia

Para esta pesquisa foi desenvolvida a metodologia D-AI^2-M — uma metodologia de Inteligência Artificial Centrada em Dados (DAI) — para selecionar as combinações mais eficientes de DAI e AIM para previsão de séries temporais de produção de etanol. Dentre as contribuições da metodologia em relação aos trabalhos relacionados destaca-se: i) O fato da metodologia não se concentrar em apenas um tipo de algorítimo AIM e ii) Avaliar quantitativamente o impacto de diferentes metodologias DAI em cenários distintos de produção de etanol combustível por unidade geográfica.

A metodologia proposta é apresentada na Figura IV.1, que descreve o processo para selecionar, treinar e validar modelos de séries temporais para previsão mensal da produção de etanol anidro e hidratado nos principais estados brasileiros. Tal diagrama detalha a segregação de dados, as fases de treinamento do modelo, a otimização de hiperparâmetros e as técnicas de validação cruzada de séries temporais utilizadas.

A Etapa A na Figura IV.1 representa a seleção das séries temporais de produção de etanol para inclusão na avaliação experimental. Esta seleção de dados considera a produção segregada por estado produtor e tipo de etanol produzido. As séries temporais são também segregadas pelo tipo de etanol uma vez que o processo e os custos de produção de cada produto variam.

A Etapa B descreve os critérios para dividir cada série temporal selecionada em subconjuntos para treinamento e validação cruzada dos modelos. Os subconjuntos de treinamento iniciais excluíram os trinta e seis meses mais recentes (últimos três anos). Três subconjuntos de validação cruzada de séries temporais foram usados, cada ciclo utilizando os 12 meses seguintes para validação, com todos os dados anteriores retribuídos como conjunto de treinamento.

O processo de previsão proposto considera a avaliação de diferentes modelos autorregressivos submetidos a diferentes pré-processamentos DAI. Este trabalho se refere a cada par *modelo-pré-processamento* como um D-AI^2-M. Um dos objetivos da metodologia é medir o desempenho dos D-AI^2-M adotando um processo de benchmarking [Salles et al., 2017]. A Etapa C1 da Figura IV.1 adota o ARIMA como modelo de referência proposto por Salles et al. [2017].

As Etapas C2, C2.1 e C2.2 cobrem o treinamento dos modelos D-AI^2-M. Os modelos autorregressivos apresentados neste artigo incluem ELM, MLP, SVR, Conv1D e LSTM. A etapa C2.1 indica que tais modelos foram treinados e avaliados separadamente usando os seguintes métodos de pré-processamento: DIFF, AN e GMM. A Etapa C2.2 detalha as estratégias de otimização de

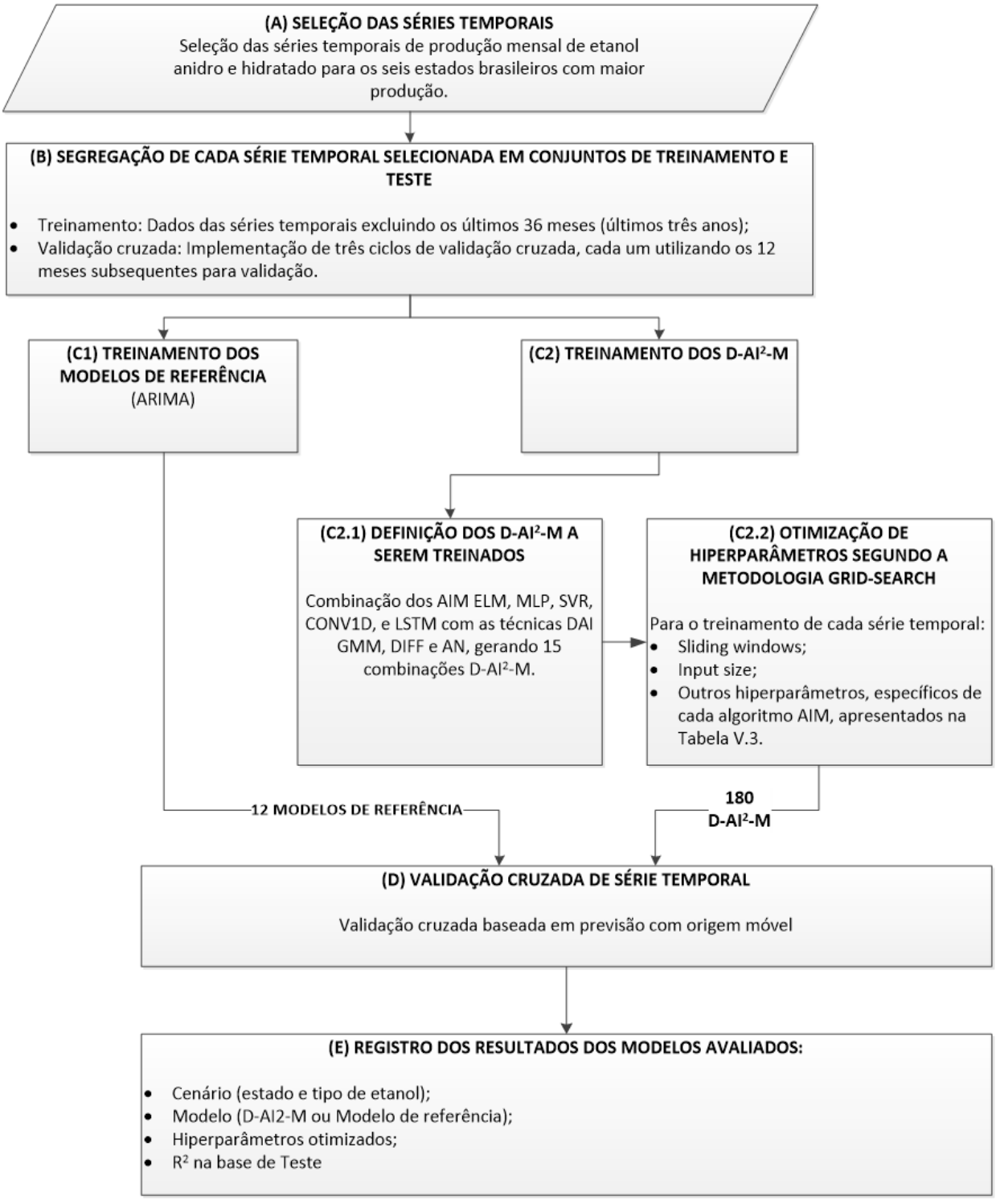


Figura IV.1: Processo D-AI^2-M - Seleção dos melhores modelos para previsão mensal brasileira da produção de etanol)

hiperparâmetros, tanto gerais (aplicáveis a todos os AIMs) quanto específicas (adaptadas a cada tipo de modelo), usadas no D-AI^2-M.

A Etapa D apresenta a metodologia de Previsão com Origem Móvel, do inglês *Rolling Forecast Origin* [Hyndman and Athanasopoulos, 2018], como o critério de validação cruzada de séries temporais aplicado aos modelos ARIMA e D-AI^2-M. Esta metodologia envolve a criação de uma série

de conjuntos de validação, onde cada subconjunto de teste se move progressivamente no tempo. Correspondentemente, cada subconjunto de treinamento inclui todos os dados anteriores ao seu respectivo subconjunto de teste. Tal abordagem é importante pois avalia a capacidade do modelo de generalizar para diferentes horizontes temporais ao simular cenários de previsão contínua (onde novos dados são constantemente incorporados ao conjunto de treino). Isso resulta em uma avaliação mais realista e robusta do desempenho do modelo avaliado.

A Etapa E da Figura IV.1 consolida a estratégia para organizar e registrar os dados relacionados ao treinamento e desempenho obtido pelos modelos treinados (benchmarking e D-AI^2-M). O desempenho dos modelos foi avaliado usando a métrica R^2 (coeficiente de determinação), que pode variar de $-\infty$ (piores resultados) até 1 (melhor resultado possível). A escolha do R^2 como métrica para avaliação dos resultados se baseou em dois critérios: i) *Interpretabilidade*: Um valor de R^2 igual a 1 indica um ajuste perfeito do modelo, onde toda a variância é explicada. Se o valor de R^2 é igual a zero isso indica que o modelo não explica nenhuma variância, o que colocaria o desempenho do modelo tão bom quando apenas prever a média dos valores. valores negativos de R^2, por sua vez, indicam que o modelo é pior que uma simples média, sugerindo que a previsão está muito distante dos valores reais e ii) *Comparabilidade*: R^2 possibilita comparar o desempenho de diferentes modelos, mesmo quando as séries temporais envolvidas estão em escalas significativamente diferentes [Chicco et al., 2021]. A fórmula para o cálculo da métrica R^2 é apresentada na Equação IV.1 onde $\hat{Y}_i$ são os valores previstos pelo modelo, Y_i são os valores reais e $\bar{y}_i$ a média dos valores reais [Chicco et al., 2021]. O desempenho dos modelos de benchmarking e D-AI^2-M é documentado separadamente para cada subconjunto de treinamento e teste, conforme definido na Etapa B. O desempenho de cada modelo é determinado considerando-se os subconjuntos de teste para cada série temporal selecionada.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n}(\bar{Y}_i - Y_i)^2} \tag{IV.1}$$

Capítulo V Avaliação experimental

Este capítulo descreve a avaliação experimental quantitativa, que foi conduzida de acordo a metodologia abordada no Capítulo IV. Este capítulo está subdividida em três seções principais: a) Conjunto de dados utilizado, b) Implementação e c) Resultados.

V.1 Conjuntos de dados

O conjunto de dados utilizado na avaliação experimental compreendeu a série histórica da produção brasileira de etanol anidro e hidratado, fornecido pela ANP [ANP, 2024a]. Esta série temporal tem frequência mensal, de janeiro de 2012 a dezembro de 2023, e inclui os seguintes atributos: i) Ano, ii) Mês, iii) Unidade da Federação, iv) Tipo de etanol, e v) Produção em m^3. Para a condução dos experimentos foram selecionados dados dos seis estados brasileiros com maior produção de etanol (SP, GO, MG, MT, MS e PR) segregados pelo tipo de etanol produzido.

A série histórica da produção de etanol anidro e hidratado para cada um dos estados selecionados é apresentada na Figura V.1. A análise da figura revela diferenças significativas na escala de volume de etanol produzido entre os estados, com São Paulo (SP) e Goiás (GO) historicamente liderando a produção. Os estados de São Paulo e Mato Grosso do Sul (MS) parecem ter sofrido uma redução de produção mais intensa no ano 2020 como efeito da pandemia de COVID-19, embora os estados de Goiás, Minas Gerais (MG) e Paraná (PR) também tenham reduzido a produção de etanol combustível nesse ano. Identificou-se um notável aumento contínuo na produção de ambos os tipos de etanol em Mato Grosso (MT) a partir de 2018, mesmo durante a pandemia de COVID-19. Essa tendência de crescimento contínuo recente não foi observada em outros estados. Tal tendência pode ser parcialmente explicada pela crescente utilização de milho para a produção de etanol. Observado-se novamente a Figura II.2 nota-se que embora haja uma queda acentuada na produção brasileira de etanol de cana de açúcar entre os anos de 2020 e 2022 (impactos da COVID-19) a produção de etanol de milho seguiu crescente, inclusive nesse período.

V.2 Implementação

Todos os experimentos foram conduzidos usando a linguagem de programação R, suportada principalmente pelo pacote DAL Toolbox [Ogasawara et al., 2023]. O repositório para esta avaliação

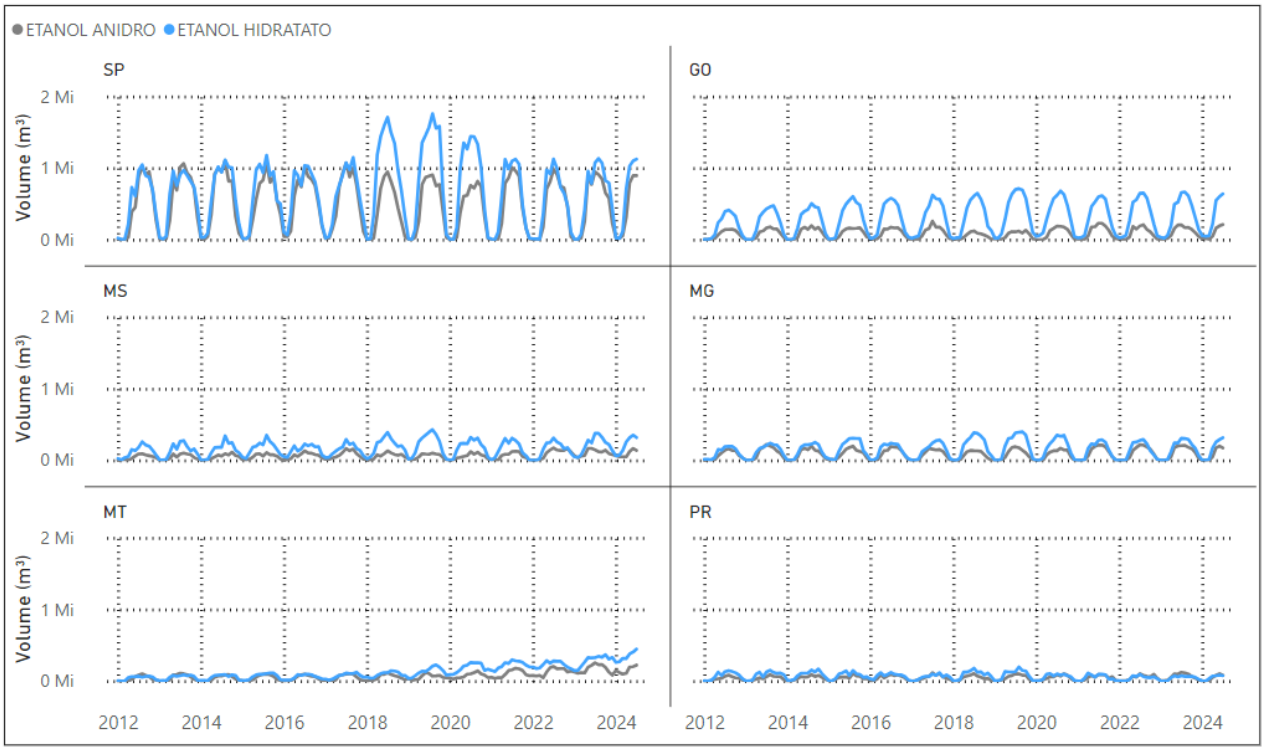


Figura V.1: Séries temporais de produção de etanol anidro e hidratado dos principais estados produtores. Fonte: Adaptado de ANP [2024a]

experimental pode ser acessado em https://github.com/cefet-rj-dal/DAI2M. A avaliação experimental utilizou modelos ARIMA ajustados automaticamente pela função $auto-arima()$ [Hyndman and Khandakar, 2008] como benchmarking. A função $auto-arima()$ faz parte do pacote *Forecast*, também desenvolvido na linguagem R [Hyndman and Athanasopoulos, 2018].

Para o treinamento D-AI^2-M, a otimização de hiperparâmetros foi realizada usando a metodologia Grid Search. Este processo de otimização utilizou intervalos de parâmetros comuns para todos os AIMs, incluindo janelas deslizantes e tamanho de entrada, bem como intervalos de parâmetros específicos adaptados para cada tipo de algoritmo AIM. Esses intervalos de hiperparâmetros são detalhados na Tabela V.1. A notação da tabela ([i: j, k]) indica uma sequência numérica começando de i a j com um tamanho de passo de k.

Tabela V.1: Hiperparâmetros para a otimização dos D-AI^2-M

D-AI^2-M	**Hiperparâmetros**
Todos os AIM	Sliding windows:[9:18,3]; Input size:[1:10,1]
ELM	nhid:[1:20,1]; actfun:[*sig*, *radbas*, *tribas*, *relu*, *purelin*]
SVR	kernel:[*radial*, *poly*, *linear*, *sigmoid*]; epsilon:[0:1,0.05]; cost:[1,10,1]
MLP	size:[1:10,1]; decay:[0,1,0.05]; maxit=700
Conv1D	epochs=700
LSTM	epochs=700

Os critérios para criação dos conjuntos de dados utilizados para validação cruzada de séries

temporais dos modelos estão detalhados na Figura V.2. Os números em cada retângulo representam o ano dos respectivos dados (conjunto de 12 meses). Para cada ciclo de validação cruzada de série temporal, a cor cinza claro indica que o retângulo faz parte do subconjunto de treinamento enquanto a cor cinza escuro indica que faz parte do subconjunto de validação.

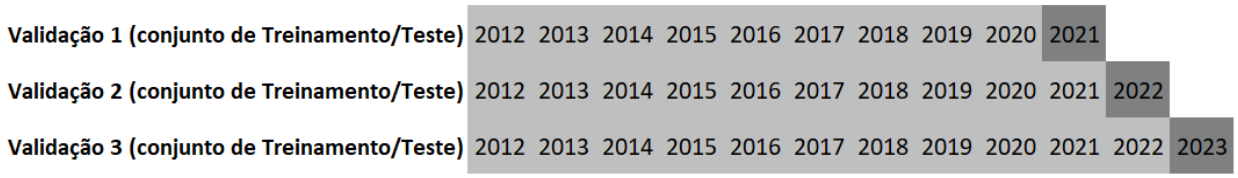


Figura V.2: Detalhes da validação cruzada de séries temporais (avaliação com origem móvel)

V.3 Resultados

Para uma melhor interpretação do ganho de desempenho obtido pelos D-AI^2-M selecionados pela metodologia proposta, apresenta-se antes i) comparações entre os desempenhos médio dos ARIMA versus os AIM por cenário avaliado (Tabela V.2) e ii) comparação entre os desempenho médios dos ARIMA versus DAI por cenário avaliado (Tabela V.3). As comparações apresentadas em tais tabelas tem o objetivo de mostrar que os modelos ARIMA ajustados automaticamente apresentam, em média, resultados melhores que os AIM ou as técnicas de DAI isoladamente (sem considerar ainda o processo de seleção das melhores combinações de DAI e AIM proposto).

Na Tabela V.2 são apresentados os resultados da comparação entre ARIMA e os diferentes AIMs. Pode-se observar que nessa abordagem os modelos de referência ARIMA apresentam melhor desempenho na maioria dos cenários (sete em doze). Para o cenário de produção de etanol hidratado no estado de Goiás (GO HID), embora o ARIMA não tenha apresentado o melhor desempenho, seu desempenho é muito próximo dos demais. É importante notar também que todos os modelos tiveram um bom desempenho nesta série temporal. O ARIMA não teve o melhor desempenho nos cenários MS ANI, MT ANI e MT HID, mas nenhum dos modelos apresentou individualmente bons resultados para essas séries temporais.

A tabela V.3 apresenta os resultados da comparação entre os modelos ARIMA e os resultados de cada pré-processamento DAI. Mais uma vez, pode-se observar que os modelos de referência ARIMA apresentam melhor desempenho na maioria dos cenários (oito em doze). Além do ARIMA, apenas o DIFF saiu vencedor em alguns cenários (GO HID, MS ANI, MT ANI e PR HID).

Analisando os R^2 médios obtidos pelos modelos ARIMA nos diferentes cenários nota-se que cinco dos seis cenários com os piores desempenhos (Valores de R^2 baixos, incluindo valores negativos) estão relacionados com estados que introduziram a produção de etanol de milho (Mato Grosso, Mato Grosso do Sul e Paraná). A exceção é O estado de Goiás que tem mantido o volume produzido de

Tabela V.2: Resultados da comparação entre ARIMA e os AIM (R^2 médio na validação cruzada de série temporal)

MODELO	GO ANI	GO HID	MG ANI	MG HID	MS ANI	MS HID
ARIMA	**0,83**	0,89	**0,89**	**0,84**	-0,34	**0,60**
CONV1D	0,76	**0,91**	0,73	0,37	-61,64	0,26
ELM	-3,69	**0,91**	-7,11	0,58	**0,41**	0,58
LSTM	0,68	0,90	0,80	0,69	0,03	0,15
MLP	0,71	0,90	0,83	-0,52	0,14	0,05
SVM	0,77	0,90	0,79	0,68	0,30	0,38
MODELO	**MT ANI**	**MT HID**	**PR ANI**	**PR HID**	**SP ANI**	**SP HID**
ARIMA	-0,77	0,02	**0,54**	-0,69	**0,85**	**0,71**
CONV1D	-1,36	-0,63	0,02	-1,18	-0,98	0,09
ELM	-0,90	-1,50	0,50	-0,06	0,77	0,45
LSTM	0,26	**0,35**	0,03	**0,11**	-2,13	-1,05
MLP	-99,6	0,34	0,39	-0,05	0,69	-0,01
SVM	**0,33**	-0,56	0,35	-0,02	0,64	0,51

Tabela V.3: Resultados da comparação entre ARIMA e os DAI (R^2 médio na validação cruzada de séries temporais)

MODELO	GO ANI	GO HID	MG ANI	MG HID	MS ANI	MS HID
ARIMA	**0,83**	0,89	**0,89**	**0,84**	-0,34	**0,60**
AN	-2,03	0,86	-3,97	-0,13	0,09	0,24
DIFF	0,79	**0,93**	0,78	0,64	**0,30**	0,25
GMM	0,78	0,92	0,80	0,58	-36,85	0,36
MODELO	**MT ANI**	**MT HID**	**PR ANI**	**PR HID**	**SP ANI**	**SP HID**
ARIMA	-0,77	**0,02**	**0,54**	-0,69	**0,85**	**0,71**
AN	-61,2	-0,84	0,27	-0,15	0,55	0,32
DIFF	**0,54**	-0,03	-0,06	**-0,13**	-1,29	-0,39
GMM	-0,06	-0,33	0,56	-0,45	0,14	0,06

etanol de milho relativamente estável nos últimos anos. A introdução do milho como matéria prima pode ter interferido nas características estatísticas das respectivas séries temporais, o que pode ter interferido no desempenho dos modelos ARIMA. Ainda assim, na comparação geral ARIMA x AIM e ARIMA x DAI o desempenho médios dos modelos ARIMA se destaca.

Entretanto, a Tabela V.4 apresenta os resultados da comparação entre os modelos ARIMA e os D-AI^2-M. Nesta comparação, modelos D-AI^2-M superam os modelos de referência ARIMA em todos os cenários avaliados. Houve apenas dois empates dos D-AI^2-M com os modelos ARIMA para os cenários ANI MG e ANI SP. Esta primeira análise ARIMA x D-AI^2-M considera os desempenhos *locais*, ou seja: específicos para cada série temporal [Montero-Manso and Hyndman, 2021] ou cenário. Os resultados locais apresentados pelos D-AI^2-M vencedores foram superiores aos melhores resultados médios identificados nas Tabelas V.2 e V.3. Ainda nessa abordagem *local* destaca-se o desempenho do D-AI^2-M *DIFF+ELM* que, além de ser a combinação vencedora em três cenários, apresentou bom desempenho em relação aos demais modelos em todos os demais.

Além do bom desempenho do *DIFF+ELM* na abordagem local, tal combinação apresentou o

Tabela V.4: Resultados da comparação Local entre ARIMA e os D-AI^2-M (R^2 médio na validação cruzada de séries temporais)

MODELO	GO ANI	GO HID	MG ANI	MG HID	MS ANI	MS HID
ARIMA	0,83	0,89	**0,89**	0,84	-0,34	0,60
AN+CONV1D	0,81	0,88	0,79	0,23	-0,54	0,11
AN+ELM	-12,8	0,87	-23,1	0,25	**0,48**	0,46
AN+LSTM	0,71	0,85	0,83	0,76	-0,08	0,42
AN+MLP	0,53	0,91	0,76	-2,48	0,16	0,24
AN+SVM	0,61	0,81	0,87	0,56	0,45	0,00
DIFF+CONV1D	0,72	0,89	0,83	0,31	0,27	0,56
DIFF+ELM	0,86	**0,95**	**0,89**	**0,87**	0,42	0,67
DIFF+LSTM	0,84	0,89	0,71	0,83	0,18	-0,22
DIFF+MLP	0,71	0,94	0,87	0,35	0,44	-0,46
DIFF+SVM	0,80	**0,95**	0,62	0,82	0,16	**0,69**
GMM+CONV1D	0,76	**0,95**	0,56	0,57	-184,7	0,12
GMM+ELM	**0,90**	0,92	0,87	0,62	0,32	0,61
GMM+LSTM	0,47	**0,95**	0,88	0,49	-0,02	0,25
GMM+MLP	0,88	0,86	0,85	0,58	-0,17	0,36
GMM+SVM	**0,90**	0,93	0,86	0,66	0,30	0,46

MODELO	MT ANI	MT HID	PR ANI	PR HID	SP ANI	SP HID
ARIMA	-0,77	0,02	0,54	-0,69	**0,85**	0,71
AN+CONV1D	-4,77	-0,16	0,29	-0,79	-0,08	0,52
AN+ELM	-2,94	-4,98	0,29	0,16	0,72	0,44
AN+LSTM	0,45	0,34	0,45	**0,34**	0,78	0,71
AN+MLP	-299,3	**0,56**	0,01	-0,13	0,74	-0,39
AN+SVM	0,35	0,05	0,31	-0,32	0,61	0,32
DIFF+CONV1D	0,57	-1,91	-0,70	-0,94	-0,45	0,49
DIFF+ELM	0,52	0,47	0,59	0,24	0,81	0,66
DIFF+LSTM	**0,62**	0,48	-0,98	0,26	-8,02	-4,14
DIFF+MLP	0,52	0,52	0,60	-0,20	0,57	0,24
DIFF+SVM	0,48	0,29	0,22	0,01	0,62	**0,80**
GMM+CONV1D	0,12	0,19	0,49	-1,79	-2,41	-0,75
GMM+ELM	-0,29	0,02	**0,63**	-0,58	0,79	0,25
GMM+LSTM	-0,28	0,24	0,62	-0,27	**0,85**	0,28
GMM+MLP	-0,03	-0,07	0,56	0,17	0,75	0,12
GMM+SVM	0,16	-2,01	0,51	0,25	0,70	0,41

melhor desempenho *global*, ou seja, o melhor desempenho médio considerando o conjunto de todas as séries temporais utilizadas na avaliação experimental [Montero-Manso and Hyndman, 2021]. Tal desempenho foi medido utilizando-se a métrica R^2 em duas abordagens: i) *Seleção*: Avaliação utilizando o ano de 2021 como base de teste e ii) *Validação*: Validação cruzada de séries temporais, considerando os anos de 2022 e 2023. A Tabela V.5 apresenta os resultados de desempenho global, que consolidam a eficácia do D-AI^2-M na modelagem mais eficiente para as previsões propostas, também nessa abordagem. para confirmar o desempenho superior do *DIFF+ELM* foi realizado um teste de Wilcoxon pareado [Wilcoxon, 1947], comparando o desempenho de tal modelo na etapa de validação cruzada com o desempenho do DIFF+SVM (modelo com o segundo melhor desempenho na etapa de seleção). Tal teste confirmou a superioridade do DIFF+ELM, com um p-valor de 6.387×10^{-5}. Além disso, uma avaliação do tamanho do efeito [Tomczak and Tomczak, 2014] foi realizada e indicou uma diferença de magnitude grande entre esses modelos (r = 0.81).

Tabela V.5: Resultados da comparação Global entre ARIMA e os D-AI^2-M (R^2 médio na avaliação e na validação cruzada de séries temporais)

D-AI^2-M	**Seleção: Avaliação ± DP (2021)**	**Validação cruzada± DP (2022 e 2023)**
DIFF+ELM	**0,78 ±0,14**	**0,60 ±0,35**
DIFF+SVM	0,63 ±0,24	0,50 ±0,43
GMM+MLP	0,61 ±0,37	0,30 ±0,59
GMM+LSTM	0,61 ±0,48	0,25 ±0,60
AN+LSTM	0,54 ±0,18	0,55 ±0,46
ARIMA	0,52 ±0,50	0,29 ±0,87
GMM+ELM	0,50 ±0,67	0,38 ±0,62
AN+SVM	0,44 ±0,28	0,36 ±0,55
GMM+CONV1D	0,36 ±0,69	-23,41 ±110,78
AN+CONV1D	0,29 ±0,56	-0,48 ±3,02
DIFF+MLP	0,25 ±0,80	0,51 ±0,32
GMM+SVM	0,15 ±1,91	0,44 ±0,37
DIFF+CONV1D	0,07 ±1,06	0,04 ±0,93
AN+MLP	-0,37 ±2,61	-37,1 ±127,23
AN+ELM	-1,17 ±3,63	-4,43 ±15,99
DIFF+LSTM	-1,85 ±7,23	-0,14 ±2,95

Capítulo VI Conclusões

Esta pesquisa indica que o uso de técnicas DAI foi crucial para melhorar o desempenho dos AIM nos cenários de avaliação. Um ponto a destacar é a importância da utilização de técnicas de validação cruzada de séries temporais para analisar a robustez dos modelos. Essa análise mais robusta aumenta a confiabilidade dos resultados obtidos e das conclusões deste trabalho.

Os resultados incentivam o uso dos AIM, especialmente quando apoiado por uma metodologia abrangente de avaliação e seleção de modelos. Isso indica que a metodologia proposta se mostrou eficaz na identificação das melhores combinações de DAI e AIM, visando selecionar os modelos mais eficientes para cada cenário (modelos com os melhores desempenhos locais). Além disso, o D-AI^2-M apoiou a identificação da combinação D-AI^2-M com melhor desempenho global para o conjunto completo dos cenários selecionados.

É importante destacar que em parte do período de treinamento dos modelos e da validação cruzada das séries temporais o planeta viveu a pandemia da COVID-19. Este evento global afetou significativamente os padrões de consumo e a produção de combustíveis, incluindo o etanol. Consequentemente, as séries temporais refletiram os impactos da pandemia em graus variados, influenciando assim os modelos preditivos. Apesar desse desafio, o D-AI^2-M se mostrou eficiente na predição da produção mensal de etanol no Brasil.

Referências

Al-Fattah, S. M. A new artificial intelligence GANNATS model predicts gasoline demand of Saudi Arabia. *Journal of Petroleum Science and Engineering*, 194, 2020.

Al-Ghandoor, A., Jaber, J., Al-Hinti, I., and Abdallat, Y. Statistical assessment and analyses of the determinants of transportation sector gasoline demand in jordan. *Transportation Research Part A: Policy and Practice*, 50:129–138, 2013.

Anh Tran, T. Comparative analysis on the fuel consumption prediction model for bulk carriers from ship launching to current states based on sea trial data and machine learning technique. *Journal of Ocean Engineering and Science*, 6(4):317 – 339, 2021.

ANP. Gasolina. url https://www.gov.br/anp/pt-br/assuntos/producao-de-derivados-de-petroleo-e-processamento-de-gas-natural/producao-de-derivados-de-petroleo-e-processamento-de-gas-natural/gasolina, 2022.

ANP. Produção de biocombustíveis. Technical report, https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos/producao-de-biocombustiveis, 2024a.

ANP. Vendas de derivados de petróleo e biocombustíveis. Technical report, https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-estatisticos/de/vdpb/, 2024b.

Azadeh, A., Behmanesh, I., Vafa Arani, H., and Sadeghi, M. H. An integrated fuzzy mathematical programming-analysis of variance approach for forecasting gasoline consumption with ambiguous inputs: Usa, canada, japan, iran and kuwait. *International Journal of Industrial and Systems Engineering*, 18(2):159–184, 2014a.

Azadeh, A., Khakestani, M., and Saberi, M. A flexible fuzzy regression algorithm for forecasting oil consumption estimation. *Energy Policy*, 37(12):5567–5579, 2009.

Azadeh, A., Manzour Alajdad, S. M. H., and Aliheidari Bioki, T. A neuro-fuzzy regression approach for estimation and optimisation of gasoline consumption. *International Journal of Services and Operations Management*, 17(2):221–256, 2014b.

Badamchizadeh, S., Latibari, A. J., Tajdini, A., Pourmousa, S., and Lashgari, A. Modeling current and future role of agricultural waste in the production of bioethanol for gasoline vehicles. *BioResources*, 16(3):4798 – 4813, 2021.

Badr, E., Nasr, G., and Dibeh, G. Econometric modeling of gasoline consumption: A cointegration analysis. *Energy Sources, Part B: Economics, Planning and Policy*, 3(3):305 – 313, 2008.

Barros, A. C., Ferreira, Pedro Guilherme Costa ans Mattos, D. M. d., Oliveira, I. C. d., and Duca, V. E. l. d. A. *Análise de Séries Temporais em R: Curso Introdutório*. FGV IBRE, 2018.

Bhanja, S. and Das, A. Deep learning-based integrated stacked model for the stock market prediction. *Int. J. Eng. Adv. Technol*, 9(1):5167–5174, 2019.

Cervero, R. Short-run forecasting of highway gasoline consumption in the united states. *Transportation Research Part A: General*, 19(4):305 – 313, 1985.

Chen, H., Tong, Y., and Wu, L. Forecast of energy consumption based on fgm (1, 1) model. *Mathematical Problems in Engineering*, 2021:1–11, 2021.

Chicco, D., Warrens, M. J., and Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7:1 – 24, 2021.

da Agricultura e Pecuária, M. Cronologia da mistura carburante etanol anidro - gasolina no brasil. url https://www.gov.br/agricultura/pt-br/assuntos/sustentabilidade/agroenergia/arquivos/cronologia-da-mistura-carburante-etanol-anidro-gasolina-no-brasil.pdf/view, 2022.

Dey, B., Roy, B., Datta, S., and Ustun, T. S. Forecasting ethanol demand in india to meet future blending targets: A comparison of arima and various regression models. *Energy Reports*, 9:411–418, 2023.

Dritsaki, C., Niklis, D., and Stamatiou, P. Oil consumption forecasting using arima models: an empirical study for greece. *International Journal of Energy Economics and Policy*, 11(4):214–224, 2021.

Figueira, S. R., Burnquist, H. L., and Bacchi, M. R. P. Forecasting fuel ethanol consumption in Brazil by time series models: 2006-2012. *Applied Economics*, 42(7):865 – 874, 2010.

Fink, R. and Medved, S. Global perspectives on first generation liquid biofuel production. *Turkish Journal of Agriculture and Forestry*, 35(5):453 – 459, 2011.

Gao, L., Lu, P., Qiao, F., Li, J. Q., Zhang, Y., and Ren, Y. Evaluating the Impact of COVID-19 on Transportation Infrastructure Funding in the United States. In *International Conference on Transportation and Development 2022: Application of Emerging Technologies - Selected Papers from the Proceedings of the International Conference on Transportation and Development 2022*, volume 6, pages 134 – 142, 2022.

Gkerekos, C., Lazakis, I., and Theotokatos, G. Machine learning models for predicting ship main engine fuel oil consumption: A comparative study. *Ocean Engineering*, 188:106282, 2019.

Gosso, A. and Martinez-de Pison, F. elmnn: Implementation of elm (extreme learning machine) algorithm for slfn (single hidden layer feedforward neural networks). *R package version*, 1, 2012.

Gujarati, D. N. *Essentials of Econometrics*. SAGE, 2021.

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

Hyndman, R. J. and Khandakar, Y. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1 – 22, 2008.

Jaber, J. O., Al-Ghandoor, A. M., Al-Hinti, I., and Sawallha, S. A. Prediction of energy consumption of passenger transportation and ghg emissions in jordan. *International Journal of Global Warming*, 4(2):90–112, 2012.

Jarrahi, M. H., Memariani, A., and Guha, S. The Principles of Data-Centric AI. *Communications of the ACM*, 66(8):84 – 92, 2023.

Jeon, H. The impact of climate change on passenger vehicle fuel consumption: Evidence from us panel data. *Energies*, 12(23):4460, 2019.

Karp, S. G., Medina, J. D. C., Letti, L. A. J., Woiciechowski, A. L., de Carvalho, J. C., Schmitt, C. C., de Oliveira Penha, R., Kumlehn, G. S., and Soccol, C. R. Bioeconomy and biofuels: the case of sugarcane ethanol in Brazil. *Biofuels, Bioproducts and Biorefining*, 15(3):899 – 912, 2021.

Kialashaki, A. and Reisel, J. Transport energy demand modeling of the united states using artificial neural networks and multiple linear regressions. In *ASME 2014 8th International Conference on Energy Sustainability, ES 2014 Collocated with the ASME 2014 12th International Conference on Fuel Cell Science, Engineering and Technology*, volume 2, 2014.

Li, H., Liu, Y., Luo, X., and Duan, H. A novel nonlinear multivariable verhulst grey prediction model: A case study of oil consumption forecasting in china. *Energy Reports*, 8:3424–3436, 2022a.

Li, Z., Rose, J. M., and Hensher, D. A. Forecasting automobile petrol demand in Australia: An evaluation of empirical models. *Transportation Research Part A: Policy and Practice*, 44(1):16 – 38, 2010.

Li, Z., Zhou, B., and Hensher, D. A. Forecasting automobile gasoline demand in Australia using machine learning-based regression. *Energy*, 239, 2022b.

Manowska, A. and Bluszcz, A. Forecasting Crude Oil Consumption in Poland Based on LSTM Recurrent Neural Network. *Energies*, 15(13), 2022.

Mardiana, S., Saragih, F., and Huseini, M. Forecasting gasoline demand in indonesia using time series. *International Journal of Energy Economics and Policy*, 10(6):132 – 145, 2020.

Marquez, J. P., De Oliveira Ribeiro, C., Santoyo, E. R., and Fernandez, V. F. Ethanol Fuel Demand Forecasting in Brazil Using a LSTM Recurrent Neural Network Approach. *IEEE Latin America Transactions*, 19(4):551 – 558, 2021.

Melikoglu, M. Demand forecast for road transportation fuels including gasoline, diesel, LPG, bio-ethanol and biodiesel for Turkey between 2013 and 2023. *Renewable Energy*, 64:164 – 171, 2014.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2023.

Montero-Manso, P. and Hyndman, R. J. Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37(4):1632 – 1653, 2021.

Mubarak, S. M. J., Crampton, A., Carter, J., and Parkinson, S. Robust data expansion for optimised modelling using adaptive neuro-fuzzy inference systems. *Expert Systems with Applications*, 189:116138, 2022.

MY, L. and GS, Z. Production trends analysis of fuel ethanol based on hubbert model. In *PROCEEDINGS OF THE 2013 INTERNATIONAL CONFERENCE ON ENERGY*. INTERNATIONAL CONFERENCE ON ENERGY, 2013.

Nasr, G., Badr, E., and Joun, C. Backpropagation neural networks for modeling gasoline consumption. *Energy Conversion and Management*, 44(6):893 – 905, 2003.

Nielsen, A. *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*. O'Reilly Media, Inc., 2019.

Niven, R. K. Ethanol in gasoline: Environmental impacts and sustainability review article. *Renewable and Sustainable Energy Reviews*, 9(6):535 – 555, 2005.

Ogasawara, E., Castro, A., Borges, H., Carvalho, D., Santos, J., Bezerra, E., and Coutinho, R. daltoolbox: Leveraging Experiment Lines to Data Analytics, 2023.

Ogasawara, E., Martinez, L. C., De Oliveira, D., Zimbrão, G., Pappa, G. L., and Mattoso, M. Adaptive Normalization: A novel data normalization approach for non-stationary time series. In *Proceedings of the International Joint Conference on Neural Networks*, 2010.

Ogasawara, E., Murta, L., Zimbrão, G., and Mattoso, M. Neural networks cartridges for data mining on time series. In *Proceedings of the International Joint Conference on Neural Networks*, pages 2302 – 2309, 2009.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, 88:105906, 2021.

Renewable Fuels Association. Annual Ethanol Production. Technical report, https://ethanolrfa.org/markets-and-statistics/annual-ethanol-production, 2024.

Rodrigues, L., Rodrigues, L., and Bacchi, M. R. P. Performance evaluation of forecasting models based on time series and machine learning techniques: an application to light fuel consumption in brazil. *International Journal of Energy Sector Management*, 16(4):636 – 658, 2022.

Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach*. Pearson, 2019.

Sadeghinezhad, E., Kazi, S. N., Badarudin, A., Togun, H., Zubir, M. N., Oon, C. S., and Gharehkhani, S. Sustainability and environmental impact of ethanol as a biofuel. *Reviews in Chemical Engineering*, 30(1):51 – 72, 2014.

Salles, R., Assis, L., Guedes, G., Bezerra, E., Porto, F., and Ogasawara, E. A framework for benchmarking machine learning methods using linear models for univariate time series prediction. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2017-May, pages 2338 – 2345, 2017.

Silva, A. L. D. and Castañeda-Ayarza, J. A. Macro-environment analysis of the corn ethanol fuel development in Brazil. *Renewable and Sustainable Energy Reviews*, 135, 2021.

Tomczak, M. and Tomczak, E. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in Sport Sciences*, 1(21):19–25, 2014.

UNICAdata. Painel de histórico da área de cultivo, moagem e produção de açúcar e etanol - brasil. Technical report, http://observatoriodacana.com.br/listagem.php?idMn=4, 2023.

Waheed Bhutto, A., Ahmed Bazmi, A., Qureshi, K., Harijan, K., Karim, S., and Shakil Ahmad, M. Forecasting the consumption of gasoline in transport sector in pakistan based on arima model. *Environmental Progress and Sustainable Energy*, 36(5):1490 – 1497, 2017.

Wilcoxon, F. Probability tables for individual comparisons by ranking methods. *Biometrics*, 3(3):119 – 122, 1947.

Wong, E., Venegas, E., and Antiporta, D. Simulating the consumption of gasoline. *Simulation*, 28(5):145–152, 1977.

Yang, Y., Chen, Y., Shi, J., Liu, M., Li, C., and Li, L. An improved grey neural network forecasting method based on genetic algorithm for oil consumption of china. *Journal of Renewable and Sustainable Energy*, 8(2), 2016.

Yu, L., Liang, S., Chen, R., and Lai, K. K. Predicting monthly biofuel production using a hybrid ensemble forecasting methodology. *International Journal of Forecasting*, 38(1):3–20, 2022.

Yu, L., Ma, Y., and Ma, M. An effective rolling decomposition-ensemble model for gasoline consumption forecasting. *Energy*, 222, 2021.

Yu, L., Zhao, Y., Tang, L., and Yang, Z. Online big data-driven oil consumption forecasting with google trends. *International Journal of Forecasting*, 35(1):213–223, 2019.