Predição da Produção de Etanol nos Estados Brasileiros

Antonio Mello[1], **Diego Carvalho**[1], **Eduardo Ogasawara**[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

antonio.mello.1@aluno.cefet-rj.br

d.carvalho@ieee.org, eogasawara@ieee.org

Abstract. *Two types of fuel ethanol are produced in Brazil: hydrated ethanol, used directly as vehicle fuel, and anhydrous ethanol, mixed with gasoline in a 27% proportion. Data from the National Agency of Petroleum, Natural Gas and Biofuels (ANP) indicated that Brazilian fuel ethanol production represented nearly 22% of the total automotive fuel consumption in the country in 2023. Six states are responsible for approximately 90% of Brazil's ethanol production, presenting a logistical challenge due to the seasonality of production and the need to transport ethanol from the production regions to the distribution networks. This study aims to model and predict the monthly production of hydrated and anhydrous ethanol in the main ethanol-producing states in Brazil. To achieve this, we employ the ARIMA method for time series forecasting. The proposed methodology presented good results for the time series that did not draw attention to significant statistical change points.*

Resumo. *Dois tipos de etanol combustível são produzidos no Brasil: etanol hidratado, usado diretamente como combustível veicular, e etanol anidro, misturado na gasolina na proporção de 27%. Dados da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) indicaram que a produção brasileira de etanol combustível representou quase 22% do consumo total de combustíveis automotivos no país em 2023. Seis estados são responsáveis por aproximadamente 90% da produção brasileira de etanol, o que apresenta um desafio logístico não só devido à sazonalidade da produção mas também pela necessidade de transporte do etanol das regiões produtoras para as redes de distribuição. Este estudo visa modelar e prever produção mensal de etanol nos principais estados produtores do Brasil. Para isso, empregamos o método ARIMA para previsão de séries temporais. A metodologia proposta apresentou bons resultados para as séries temporais que não apresentaram indícios de pontos de mudança estatísticos significativos.*

1. Introdução

O etanol combustível é comercializado no Brasil em duas formas: etanol hidratado, utilizado diretamente como combustível para veículos, e etanol anidro, que atualmente é misturado na proporção de 27% na gasolina comum [ANP, 2023b]. Em 2023, o volume total de etanol comercializado no Brasil ultrapassou 28 milhões de metros cúbicos, o que representou quase 22% do volume total de combustíveis líquidos vendidos [ANP, 2023b], destacando a relevância deste biocombustível no país.

São Paulo (SP), Goiás (GO), Minas Gerais (MG), Mato Grosso (MT), Mato Grosso do Sul (MS) e Paraná (PR) se destacam na produção de etanol, representando mais de

90% do total produzido no Brasil [ANP, 2023a]. A produção nacional é majoritariamente baseada na cana-de-açúcar, embora o etanol de milho tenha ganhado relevância, especialmente em Mato Grosso e Mato Grosso do Sul [da Silva and Castañeda-Ayarza, 2021]. A produção de etanol, por ser agrícola, apresenta forte sazonalidade, conforme mostrado na Figura 1, baseada em dados da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) [ANP, 2023a].

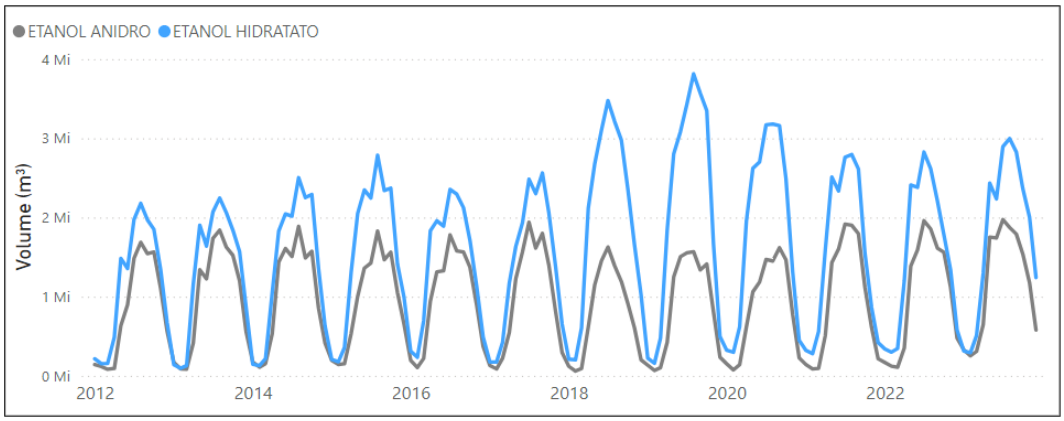

Figura 1. Produção mensal brasileira de etanol anidro e hidratado entre 2013 e 2022 [ANP, 2023a]

A rede de distribuição e revenda de etanol é complexa e muitas vezes não se relaciona diretamente apenas com os estados produtores e consumidores. Mesmo assim, a Tabela 1 fornece uma comparação ilustrativa dos volumes de etanol produzidos e consumidos, segregados pelos principais estados produtores. Essa comparação exibe o volume total de etanol produzido e consumido por estado entre 2019 e 2023. O volume de etanol anidro consumido foi estimado em 27% da gasolina consumida durante o período. Analisando os números apresentados, podemos notar o significativo volume excedente de etanol produzido em relação ao consumido para os quatro maiores estados produtores (SP, GO, MT e MS).

Tabela 1. Produção e consumo de etanol pelos seis principais estados produtores brasileiros entre 2019 e 2023 (dados da ANP [ANP, 2023a,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

Este trabalho visa apresentar uma metodologia baseada no uso do algoritmo ARIMA, técnica estatística amplamente utilizada para previsão de séries temporais, para prever a produção mensal brasileira de etanol. Para tanto esse estudo se concentra nas séries temporais de produção de etanol dos seis maiores estados produtores, segregadas por produto (anidro e hidratado).

O restante deste artigo está organizado em mais seis seções. A Seção 2 apresenta uma revisão teórica dos principais conceitos e técnicas aplicáveis à previsão da produção de etanol. A Seção 3 aprofunda-se no corpo de trabalhos existentes na área. A Seção 4 apresenta a metodologia seguida neste artigo. A Seção 5 descreve as etapas da avaliação experimental, enquanto a Seção 6 examina e discute os resultados obtidos. Finalmente, a Seção 7 apresenta as conclusões do trabalho.

2. Fundamentos

Esta seção apresenta uma revisão teórica envolvendo os principais conceitos e métodos estatísticos associados com a modelagem e predição de séries temporais envolvendo algoritmos ARIMA.

Séries Temporais. As séries temporais são sequências de dados ordenadas e coletadas em intervalos de tempo regulares e conhecidos [Al-Fattah, 2020]. Elas podem ter um ou mais dos seguintes componentes: (i) tendência, (ii) sazonalidade e (iii) ciclo. A tendência é o comportamento observado quando a série temporal apresenta um padrão crescente ou decrescente ao longo do tempo, não necessariamente linear. A sazonalidade é o padrão de repetição que ocorre em uma periodicidade específica. Ciclos 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 assim da sazonalidade [Hyndman and Athanasopoulos, 2018].

Modelos autorregressivos. Os modelos autorregressivos 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 autorregressivo), 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 utilização de modelos autorregressivos considera que os conjuntos de dados a ser utilizado é uma série temporal. Um dos modelos autorregressivos mais frequentemente utilizados é o ARIMA, apresentado no próximo item.

Modelos ARIMA. O ARIMA é um acrônimo para *Auto Regressive Integrated Moving Average* (Média Móvel Integrada Autorregressiva). Este modelo combina três componentes principais: (i) autorregressão, (ii) integração, (iii) média-móvel.

A autorregressão (AR) assume que o valor atual de uma série temporal depende dos valores defasados anteriores da série. O valor do termo p na representação do modelo ARIMA(p, d, q) indica quantos períodos anteriores estão sendo considerados. A integração (I) representa a diferenciação 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. Finalmente, a média móvel (MA) considera a média dos erros das previsões anteriores para modelar o comportamento da previsão atual. 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 pela Equação 1 [Nielsen, 2019; Hyndman and Athanasopoulos, 2018]. Nela, y_t^d é a série temporal diferenciada conforme a ordem d [Hyndman and Athanasopoulos, 2018]; ϕ_0 é uma constante também chamada de média não nula; $\sum_{i=1}^{p}(\phi_i \times y_{t-i}^d)$ é o componente Autorregressivo (AR), de ordem p; $\sum_{i=1}^{q}(\theta_i \times e_{t-i})$ é o componente de Média Móvel (MA), de ordem q; e e_t é a parte aleatória da série temporal.

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

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]. Entretanto, existem métodos para ajustes automatizados do modelo ARIMA [Nielsen, 2019], sendo o *auto-arima*, presente no pacote *forecast* desenvolvido na linguagem R [Hyndman and Athanasopoulos, 2018], o mais conhecido.

3. Trabalhos Relacionados

A previsão da produção de etanol por unidade geográfica (país, estado ou região) é um tema de pesquisa importante, considerando seu papel crucial como biocombustível. Diversos estudos têm explorado diferentes abordagens para prever a produção e o consumo de etanol combustível, utilizando métodos estatísticos e de aprendizado de máquina.

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 matérias-primas para a produção de etanol e biodiesel. Eles concluíram que um aumento na temperatura tem um impacto negativo significativo na produção de matérias-primas para etanol, enquanto a precipitação não teria impacto significativo na produção de biocombustíveis.

No Irã, Badamchizadeh et al. [2021] desenvolveram um modelo autorregressivo baseado em uma rede neural artificial para prever a necessidade anual potencial de produção de etanol, considerando três cenários diferentes de mistura de etanol na gasolina. A previsão foi baseada na demanda de gasolina, e o estudo apresentou opções viáveis para fontes de matéria-prima para etanol a partir de resíduos agrícolas.

Nos EUA, Yu et al. [2022] propuseram uma combinação de métodos para prever a produção mensal de biocombustíveis. A metodologia envolveu a aplicação da Decomposição Empírica de Modo para decompor os dados originais, seguida pelo uso da técnica LSTM para prever o componente de alta frequência e da técnica ELM para prever o componente de baixa frequência. Os resultados individuais de previsão foram integrados em um resultado agregado. A abordagem combinada apresentou melhores resultados nas métricas RMSE e MAPE quando comparada a diferentes modelos de previsão.

Na Índia, Dey et al. [2023] descobriram que o modelo ARIMA foi o mais preciso para prever a demanda anual de gasolina. Essa previsão foi utilizada para estimar a demanda de etanol, considerando a meta planejada de mistura de 20% de etanol na gasolina.

No Brasil, Figueira et al. [2010] utilizaram o modelo SARIMA (uma extensão do modelo ARIMA) para prever o consumo de etanol hidratado com base na série histórica

anual de consumo de biocombustíveis. Para prever o consumo de etanol anidro, utilizaram a função de transferência, incorporando variáveis exógenas como o PIB per capita anual brasileiro. A avaliação da correlação cruzada entre a variável exógena (PIB) e a variável dependente (consumo de gasolina) foi fundamental na seleção do melhor modelo.

Embora vários estudos tenham explorado diferentes métodos para prever a produção e o consumo de etanol, ainda há espaço para estudar o modelo ARIMA no contexto dos estados brasileiros. Este estudo visa preencher essa lacuna, investigando a eficácia deste modelo na previsão da produção de etanol nos principais estados produtores do Brasil.

4. Metodologia

A metodologia proposta visa garantir a confiabilidade dos resultados obtidos por meio da aplicação de uma sequência de práticas recomendadas a cada cenário avaliado. No contexto desse trabalho, cada cenário equivale a um par *estado produtor e tipo de etanol produzido*. O processo metodológico é dividido em quatro etapas, conforme ilustrado na Figura 2.

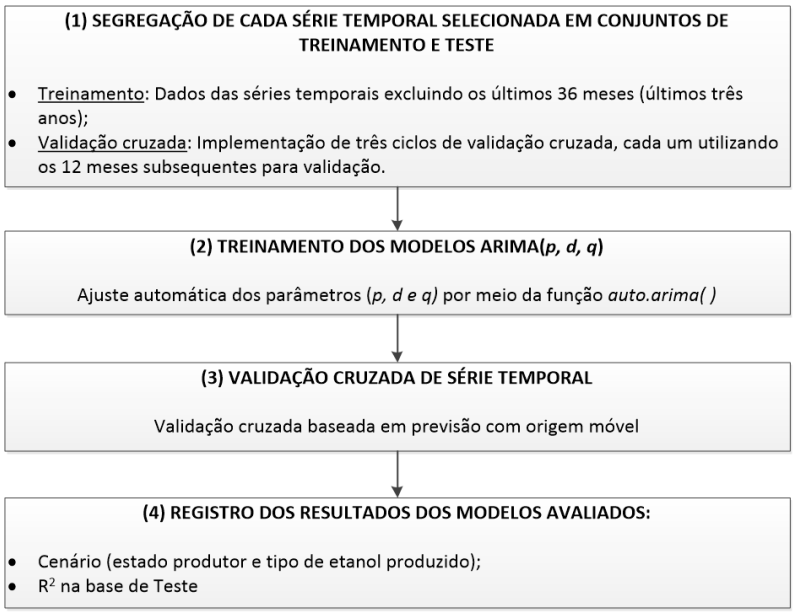

Figura 2. Etapas realizadas no treinamento e avaliação dos modelos preditivos ARIMA

A etapa (1) da metodologia descreve os critérios para a criação de uma sequência de conjuntos de treinamento e teste a partir da série temporal original. A metodologia prevê utilizar os últimos trinta e seis meses de cada série temporal para a criação de três subconjuntos de dados para a posterior validação cruzada. O processo de validação cruzada é detalhado na etapa (3).

A etapa (2) define que o treinamento dos modelos ARIMA considerar o ajuste automático de seus parâmetros pelo método *auto-arima*, garantindo a seleção otimizada dos parâmetros para cada treinamento realizado.

Em seguida, a etapa (3) orienta a utilização da metodologia de Previsão com Origem Móvel, do inglês *Rolling Forecast Origin* [Hyndman and Athanasopoulos, 2018],

como o critério de validação cruzada de séries temporais. Esta metodologia envolve a criação de uma série de conjuntos de validação, onde cada subconjunto de teste se move progressivamente no tempo.

Por fim, a etapa (4) define o registro dos resultados para cada cenário. Nessa etapa definiu-se que o desempenho dos modelos deve ser avaliado usando a métrica R^2 (coeficiente de determinação), que assume valores mais altos (próximos de um) para modelos de melhor desempenho e valores mais baixos (incluindo negativos) para modelos de pior desempenho [Chicco et al., 2021].

5. Avaliação Experimental

Todos os experimentos foram realizados utilizando a linguagem de programação R. Os experimentos foram suportados principalmente pelo pacote DAL Toolbox [Ogasawara et al., 2023].

Os dados utilizados para a avaliação experimental compreenderam as séries temporais da produção brasileira de etanol, disponibilizadas pela ANP [ANP, 2023a]. Tais séries temporais possuem frequência mensal, iniciando-se em janeiro de 2012, e incluem atributos indicativos do estado produtor e tipo de etanol produzido. As séries temporais dos seis principais estados produtores foram segregadas conforme definido na etapa (1) da metodologia proposta.

Os modelos ARIMA foram ajustados automaticamente, conforme definido na epata (2) da metodologia, pela função auto-arima() [Hyndman and Khandakar, 2008]. Tal função é parte do pacote *forecast* disponibilizado na linguagem R [Hyndman and Athanasopoulos, 2018].

Na Figura 3 é detalhado o processo de validação cruzada de séries temporais. Os números em cada retângulo representam o ano dos respectivos dados (conjuntos de 12 meses). Para cada um dos três ciclos de validação, a cor cinza-claro indica dados do subconjunto de treinamento, enquanto a cor cinza-escuro indica dados do subconjunto de validação.

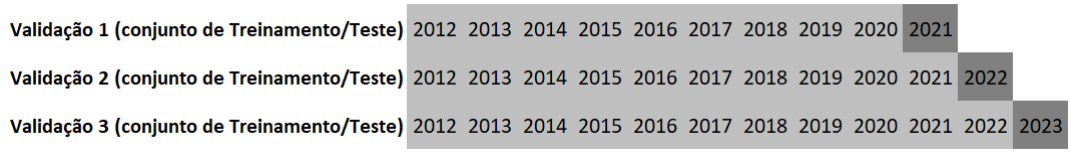

Figura 3. Validação cruzada de séries temporais (avaliação com origem móvel)

6. Resultados

Esta seção apresenta os desempenhos obtidos pelos modelos ARIMA, treinados e avaliados conforme a metodologia proposta. Na Tabela 2 são exibidos os valores de R^2 obtidos para os três ciclos de validação, por cenário selecionado. Além disso, na tabela também são exibidos os parâmetros ótimos automaticamente atribuídos a cada modelo assim como o R^2 médio alcançado para cada cenário. A Tabela 2 está organizada de forma decrescente com base nos valores de R^2 Médio $\pm$ desvios padrão.

Tabela 2. Avaliação dos modelos ARIMA

CENÁRIO	2021 R^2	2021 ARIMA	2022 R^2	2022 ARIMA	2023 R^2	2023 ARIMA	R^2 Médio $\pm$ DP
MG-ANI	0,90	(5,0,0)	0,92	(5,0,0)	0,84	(5,0,0)	0,89 $\pm$0,04
GO-HID	0,78	(2,0,2)	0,94	(2,0,2)	0,94	(2,0,2)	0,89 $\pm$0,08
SP-ANI	0,91	(4,0,1)	0,74	(4,0,1)	0,91	(4,0,1)	0,85 $\pm$0,09
MG-HID	0,78	(2,0,2)	0,92	(2,0,2)	0,82	(2,0,2)	0,84 $\pm$0,06
GO-ANI	0,79	(5,0,0)	0,84	(5,0,0)	0,85	(5,0,0)	0,83 $\pm$0,03
SP-HID	0,76	(2,0,1)	0,60	(2,0,1)	0,76	(2,0,1)	0,71 $\pm$0,08
MS-HID	0,60	(2,0,2)	0,79	(5,0,0)	0,41	(5,0,0)	0,60 $\pm$0,18
PR-ANI	0,80	(5,0,1)	0,66	(5,0,1)	0,16	(5,0,1)	0,54 $\pm$0,33
MT-HID	-0,70	(0,1,1)	0,58	(4,1,1)	0,17	(5,1,1)	0,02 $\pm$0,54
MS-ANI	0,74	(4,0,1)	0,18	(4,0,1)	-1,95	(2,0,2)	-0,34 $\pm$1,41
PR-HID	0,19	(5,0,1)	-1,37	(5,0,1)	-0,89	(5,0,0)	-0,69 $\pm$0,66
MT-ANI	-0,34	(2,0,2)	-1,85	(1,1,0)	-0,12	(1,1,0)	-0,77 $\pm$0,83

Destaca-se que nos seis cenários com maior R^2 médio (melhor desempenho) os parâmetros dos respectivos modelos ARIMA não se alteraram ao longo dos três ciclos de validação. Isso pode indicar que não houve um ponto de mudança significativo nas estatísticas das respectivas séries temporais. Nota-se ainda os menores valores de desvio padrão, indicando estabilidade nas predições em tais cenários.

Já em cinco dos seis cenários com os menores valores de R^2 médio observa-se ao menos uma variação nos parâmetros dos respectivos modelos ARIMA. Isso pode indicar a ocorrência de um ponto de mudança em tais séries temporais, o que pode ainda justificar o desempenho inferior dos modelos ARIMA e o maior desvio padrão percebido em tais cenários. Vale observar que os seis cenários de menor valor de R^2 médio envolvem estados que iniciaram, mesmo que em pequena escala, a produção de etanol de milho. A utilização do milho pode justificar os pontos de mudança nas respectivas séries temporais.

7. Conclusão

A metodologia proposta apresentou bons resultados para as séries temporais que não apresentaram indícios de pontos de mudança estatísticos significativos. Vale ressaltar que tais resultados foram obtidos a um baixo custo computacional exigido pela modelagem ARIMA. Outro ponto a destacar é a possibilidade de utilização da modelagem ARIMA como ferramenta de apoio na identificação de pontos de mudança em séries temporais.

Destacamos que, durante parte dos cinco ciclos de treinamento e teste (2020 até 2022), o planeta enfrentou a pandemia de COVID-19. Sendo assim, as séries temporais utilizadas refletiram os impactos da pandemia, influenciando os modelos preditivos.

Por fim, os resultados deste estudo têm implicações práticas importantes para o planejamento logístico e a gestão da cadeia de suprimentos de etanol no Brasil. A melhoria na previsão da produção de etanol pode ajudar a otimizar a distribuição, reduzir custos e aumentar a eficiência operacional.

Trabalhos futuros podem incluir modelos preditivos mais avançados, baseados em aprendizado de máquina. A avaliação utilizando modelos mais complexos pode identificar alternativas com melhor desempenho preditivo, principalmente em séries temporais que apresentem pontos de mudança nas características estatísticas.

Agradecimentos

Os autores agradecem ao CNPq, CAPES e FAPERJ pelo apoio parcial a esta pesquisa.

Referências

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

ANP (2023a). Production of biofuels. Technical report, https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos/producao-de-biocombustiveis.

ANP (2023b). Sales of petroleum derivatives and biofuels. Technical report, https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-estatisticos/de/vdpb/.

Badamchizadeh, S., Latibari, A. J., Tajdini, A., Pourmousa, S., and Lashgari, A. (2021). Modeling Current and Future Role of Agricultural Waste in the Production of Bioethanol for Gasoline Vehicles. *BioResources*, 16(3):4798 – 4813.

Chicco, D., Warrens, M. J., and Jurman, G. (2021). 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.

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

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

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

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

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

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

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

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

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

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

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