D-AI^2-M: Ethanol Production Forecasting in Brazil Using Data-Centric Artificial Intelligence Methodology

Antonio Mello, Lucas Giusti, Tarsila Tavares, Fernando Alexandrino, Gustavo Guedes, Jorge Soares, Rafael Barbastefano, Fabio Porto, Diego Carvalho, and Eduardo Ogasawara

***Abstract*—Ethanol is as one of Brazil's primary biofuels, with two main types: i) hydrous ethanol, used directly as vehicle fuel, and ii) anhydrous ethanol, blended at 27% into regular gasoline. In 2023, data from the National Agency of Petroleum, Natural Gas, and Biofuels (ANP) indicated that over 28 million cubic meters (m^3) of ethanol were sold in Brazil, making up nearly 22% of the total volume of liquid fuels sold in the country. Just six states account for approximately 90% of Brazilian ethanol production. The logistical challenge arises from seasonal production and the need to transport ethanol from production sites to distribution networks. Typically, econometric models like ARIMA are used for that prediction. However, with advancements in Artificial Intelligence models (AIM), the question arises: Can AIM improve monthly ethanol production predictions for Brazil's key producing states? How should data be prepared for this? This study aims to contribute to logistical planning by employing D-AI^2-M —a Data-Centric Artificial Intelligence (DAI) methodology—to aid in selecting AIM for ethanol production time series in the principal Brazilian-producing states. Our quantitative evaluation shows D-AI^2-M's superior forecasting performance in two approaches: i) Local: where different D-AI^2-M models outperform benchmarks depending on the time series, and ii) Global: where a single D-AI^2-M achieves the best mean performance across all evaluated series.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9079

***Index Terms*—Ethanol production, Data-Centric Artificial Intelligence, Time Series Forecasting.**

I. Introduction

The significance of ethanol as a fuel for light vehicles in Brazil is quite substantial. Ethanol is marketed in Brazil as hydrous ethanol, used directly as vehicle fuel, or as anhydrous ethanol, which is currently blended in a proportion of 27% in regular gasoline [1]. In 2023, the total volume of ethanol marketed in Brazil corresponded to almost 22% of the total volume of fuels sold in the country [1], illustrating the importance of this biofuel.

In addition to Brazil, several other countries produce ethanol as a vehicle fuel to be mixed with gasoline, with the blending percentages varying according to the country or region. Among these countries are the USA, members of the European Union, India, China, Canada, Thailand, and Argentina [2]. The primary motivations for utilizing ethanol blended with gasoline include reducing dependence on oil and lowering greenhouse gas emissions, particularly with higher ethanol concentrations in the mixture [3, 4].

Brazilian ethanol production is predominantly based on using sugarcane as raw material. However, ethanol production from corn has recently intensified [5]. Because ethanol production uses agricultural raw materials, it has a strong seasonal component. The monthly variation in Brazilian ethanol production is presented in Fig. 1, prepared from data provided by the National Agency of Petroleum, Natural Gas and Biofuels (ANP) [6].

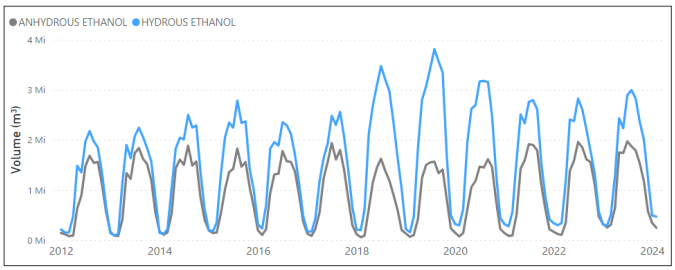

Fig. 1. Brazilian monthly production of anhydrous and hydrous ethanol between 2012 and 2023 [6].

The Brazilian states of São Paulo (SP), Goiás (GO), Minas Gerais (MG), Mato Grosso (MT), Mato Grosso do Sul (MS), and Paraná (PR) are prominent in ethanol production volume, collectively accounting for over 90% of Brazil's production [6]. The ethanol distribution and resale network is complex. It often does not only directly relate to the producing and consuming states. Even so, Table I provides an illustrative comparison, based on the period between 2019 and 2023, of the volumes of ethanol produced and consumed by state. The volume of anhydrous ethanol consumed was estimated

The associate editor coordinating the review of this manuscript and approving it for publication was Bruno Henrique Groenner Barbosa (*Corresponding author: Antonio Mello*).

A. Mello, L. Giusti, T. Tavares, G. Guedes, J. Soares, R. Barbastefano, D. Carvalho, and E. Ogasawara are with Federal Center for Technological Education of Rio de Janeiro, Rio de Janeiro, Brazil (e-mails: antonio.mello.1@aluno.cefet-rj.br, lucas.giusti@aluno.cefet-rj.br, tarsila.tavares@aluno.cefet-rj.br, gustavo.guedes@cefet-rj.br, jorge.soares@cefet-rj.br, rafael.barbastefano@cefet-rj.br, diego.carvalho@cefet-rj.br, and eduardo.ogasawara@cefet-rj.br).

F. Alexandrino is with Federal Institute of Technological Education of São Paulo, São Paulo, Brazil (e-mail: fernando.alexandrino@ifsp.edu.br).

F. Porto is with National Laboratory for Scientific Computing, Petrópolis, Brazil (e-mail: fporto@lncc.br).

to be 27% of the gasoline consumed during the period. Analyzing Table I, we can observe a significant surplus of ethanol production compared to consumption in the two largest producing states (SP and GO), as well as in MT and MS.

TABLE I
ETHANOL PRODUCTION AND CONSUMPTION BY BRAZILIAN STATE BETWEEN YEARS 2019 AND 2023 (DATA FROM ANP [6, 1])

Federation Unit	**Production (x1000 m^3)**	**Consumption (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

Ethanol production is influenced by several factors beyond seasonality, including (1) climatic conditions impacting crop productivity, (2) ethanol and sugar prices affecting production priorities, and (3) oil prices influencing ethanol's competitiveness with gasoline [7]. The COVID-19 pandemic also significantly reduced fuel demand, including ethanol [8]. Despite these influences, this study relies on the autoregressive assumption that such information is already incorporated into autoregressive terms [9], which means that models are built solely on monthly ethanol production time series, without using exogenous variables. Furthermore, while some regions in Brazil are more suitable for growing sugarcane or corn, and cultivation areas vary by state, this work does not aim to analyze spatial relationships in ethanol production across states. In this work, such an assumption is already incorporated into the autoregressive terms of the time series [9].

Given the inherent seasonality of production, variations in production scale among states, the necessity for storing and transporting ethanol from production hubs to end consumers, and the influence of climatic and market factors, the prediction of ethanol is commonly supported by econometric models, such as ARIMA. The ARIMA algorithm optimizes the model by considering the entire time series used in its training, focusing on linear relationships between observations. In contrast, Artificial Intelligence Models (AIM) can improve forecasting performance by employing advanced techniques which enable the capture of complex patterns and nonlinear relationships within the data. Furthermore, AIM can improve forecasting performance by using methods such as sliding windows to process time series by intervals rather than considering the time series as a whole.

The research problem can be articulated as follows: How can AIM be introduced to support predictions that outperform state-of-the-art models such as ARIMA? Furthermore, can Data-Centric Artificial Intelligence (DAI) initiatives [10] enable techniques to support such a goal? This study aims to explore D-AI^2-M, a DAI methodology, to build AIM to improve forecast accuracy and operational efficiency [10]. This work seeks to analyze the results obtained from two different perspectives: i) *local* performance analysis, focused on identifying the best model for each time series individually, and ii) *global* performance analysis, which aims to identify the best model considering the complete set of all time series used in the experimental evaluation [11].

Little emphasis has been given to works concerning the importance of data preprocessing for aiding the performance of predictive models. This paper addresses this gap by investigating integrating DAI techniques with autoregressive AIM. Specifically, this approach is applied to the analysis of monthly ethanol production time series in the major ethanol-producing states of Brazil. The objective is to identify the combinations of data preprocessing techniques and AIM that yield the best performance in each evaluated scenario.

The remainder of this paper is organized into six more sections. Section II provides a theoretical review of the main concepts and techniques applicable to ethanol production prediction. Section III delves into the existing body of work in the field. Section IV presents the methodology followed in this paper. Section V records the steps of the experimental evaluation, while Section VI examines and discusses the obtained results. Finally, Section VII presents the conclusions of this paper.

II. BACKGROUND

This section presents a theoretical review involving the main concepts, statistical methods, and AIM applicable in predicting the production or demand of fuel ethanol.

A. Time Series

Time series are data sequences ordered and collected at regular and known time intervals [12]. Time series can have one or more of the following components: i) Trend: Observed behavior when the time series shows an increasing or decreasing behavior over time, not necessarily linear; ii) Seasonality: A pattern of repetition noticed within a certain periodicity and iii) Cycle: Cyclical patterns that occur as fluctuations around the trend due to economic, environmental, or other factors. These cycles are not of a fixed duration and can vary in length, differing from seasonality, with a fixed and known frequency [13].

A time series can be decomposed using Equation 1. In such a case, y_t represents a time series composed by the combination of the trend component β_t and seasonality s_t.

Besides, e_t represents the part not captured, termed error or residual.

$$y_t = \beta_t + s_t + e_t \quad (1)$$

B. DAI

DAI is a concept that emphasizes the importance of handling and improving the input data for AIM. This approach offers an effective alternative for performance improvement, complementary to the refinement of algorithms [10].

An important DAI preprocessing to support AIM is transforming a time series into a sliding window (SW). Instead of considering the entire data vector, this approach divides it into SW segments of length p. Let a time series X be a sequence of n observations $< x_1, x_2, x_3, \ldots, x_n >$, where x_1 represents the value assumed by the series in the oldest observation and x_n represents the value of the series in the most recent observation. A sub-sequence of size p obtained from X that ends at position i can be represented by $seq_{i,p}(X)$ which is equivalent to the sequence of values $< x_{i-p+1}, x_{i-p+2}, \ldots, x_i >$, where $|seq_{i,p}| = p$ and $p \leq i \leq |X|$. SW is an approach to explore all possible sub-sequences of size p of a time series [14].

Another important DAI preprocessing is data normalization. Combined with SW, normalization might significantly improve AIM [15]. There are many preprocessing methods available, such as global min-max (GMM) [16], differentiation (DIFF) [9], and adaptive normalization (AN) [14].

GMM normalizes the values of a given vector of variables based on their current maximum and minimum values and the new maximum and minimum values after preprocessing. This process is illustrated in Equation 2 where x represents the original data vector, x' the normalized vector, new_{max} represents the new maximum value, and new_{min} the new minimum value of the normalized vector [14].

$$x' = (new_{max} - new_{min}) \times \frac{x - min_x}{max_x - min_x} + new_{min} \quad (2)$$

DIFF is a method used to transform a non-stationary time series into a stationary series. This is done by calculating the differences between consecutive values in the time series. This process can be repeated to obtain the second difference, the third difference, and so on until the series becomes stationary. If stationarity is required, this technique is applied before model training to ensure the input data are appropriate for the model in question, improving the prediction quality.

AN is a preprocessing method for non-stationary heteroscedastic time series. Unlike traditional sliding window techniques, AN transforms time series into data sequences where global statistical properties derived from a sample set are integrated to support the normalization. This enables AN to represent varying volatilities within its sliding windows effectively. The methodology of AN is structured into three distinct stages: (i) transforming the non-stationary time series into a stationary sequence through the creation of disjoint sliding windows, (ii) the removal of outliers, and (iii) the actual normalization. This approach offers a comprehensive solution for handling the complexities inherent in non-stationary heteroscedastic time series [14].

C. AIM

AIM encompasses a broad range of data-driven models. It includes, among others, Multi-layer Perceptron (MLP), Extreme Learn Machine (ELM), Long Short-Term Memory (LSTM), 1D Convolutional Neural Network (Conv1D), and Support Vector Regression (SVR). The MLPs are computational models inspired by the structure of brains, consisting of interconnected processing units (called *neurons*), which work together to analyze data and make predictions. Each neuron applies a function, usually nonlinear, to the weighted sum of the received signals to determine its output signal [17]. This function is referred to as the activation function. The most common activation functions are the sigmoid, ReLu, and softplus [18].

During the model training, the process where signals flow from the input layer, passing through intermediate layers until reaching the output layer, is known as *feed-forward*. At the end of each *feed-forward* cycle, the algorithm calculates and performs the back-propagation of the error (the difference between the predicted outputs and the actual outputs) and adjusts the weights of the connections between neurons. This process is repeated, usually supported by an optimization algorithm such as Gradient Descent, until the ideal weight values are found, and the model is considered trained [19, 17].

There is a type of single hidden layer artificial neural network called Extreme Learning Machine (ELM) characterized by being trained more quickly than a conventional artificial neural network. The main feature of the ELM is that the neurons' weights in the hidden layer are initialized randomly and remain fixed. Only the output layer weights are adjusted during the model training, which favors the training speed [20].

Deep learning models utilize artificial neural networks [21]. Unlike conventional neural networks, deep learning models contain multiple hidden layers, enabling them to learn complex patterns. The Long Short-Term Memory (LSTM) is a deep learning network specifically for time series processing. It includes memory cells that retain important information for longer periods. These cells use a decision-making system to selectively *remember* or *forget* information based on its relevance. The cells have three *gates*: i) the input gate for receiving data; ii) the forget gate, which manages information retention; and iii) the output gate, which controls information transmission [22].

Another deep learning technique with applications in time series is Conv1D, which incorporates the technique of convolution in data processing. In this context, convolution applies a filter (or kernel) to extract important features from a data sequence over time. This filter slides over the time series, which can capture specific characteristics and more complex patterns more accurately [23].

The SVR method is a regression-focused extension of the Support Vector Machine (SVM). The main advantage of SVR over other regression methods is the adoption of the principle of structural risk minimization, which aims to minimize an upper bound of generalization error rather than training error, thereby favoring the model's generalization [24]. SVR seeks to find a hyperplane that has the maximum number of output

y points around it within a pre-determined margin ϵ (called the ϵ-tube) while minimizing the magnitude of the weights w (to reduce model complexity). For SVR to find the separation hyperplanes, the algorithm uses functions called kernels. These functions typically map the input space to a higher-dimensional form to enhance separability of data that is not linearly separable in its original space. Common kernels used in SVR are linear, polynomial, radial, and sigmoid [25].

III. Related Work

This section is divided into two subsections: a) Works related to the prediction of ethanol or other biofuel production, and b) Works related to the prediction of gasoline or ethanol fuel consumption.

A. Prediction of Ethanol or other Biofuel Production

Fink and Medved [26] proposed using mathematical models to estimate how air temperature can affect the production of crops that serve as raw materials for ethanol and biodiesel production. An increase in temperature significantly negatively impacts the production of raw materials for ethanol and a substantial portion of the raw materials used in bio-diesel production. They proposed an equation to calculate the amount of fuel obtainable from sugarcane. The same study concluded that precipitation effects would not significantly impact biofuel production.

Badamchizadeh et al. [27] proposed an autoregressive model based on an artificial neural network with two layers (6 and 3 neurons) to predict a potential annual need for ethanol production in Iran, considering three different scenarios of ethanol blending in gasoline. They observed that such a mixture does not yet occur in that country. An indirect prediction of ethanol demand was made based on gasoline demand. Then, some potentially viable options for sources of ethanol feedstock from agricultural waste in Iran were presented.

Yu et al. [28] proposes a method for predicting monthly biofuel production in the USA, involving four steps: i) Empirical Mode Decomposition to break down the data into simpler components; ii) LSTM to predict the high-frequency component; iii) ELM to predict the low-frequency component; iv) Integration of these predictions into a single output through simple addition. The combined approach was compared with various prediction models, achieving the best results based on RMSE (Root Mean-Square Error) and MAPE (Mean Absolute Percentage Error) metrics.

B. Prediction of Gasoline or Ethanol Fuel Consumption

Melikoglu [29] proposed using semi-empirical models to predict, among others, the demand for gasoline and fuel ethanol in Turkey. Turkey's annual gasoline consumption time series was used to make both predictions (gasoline demand and ethanol to be added to gasoline). The predicted percentage of this biofuel blend in the country's gasoline was considered for ethanol consumption prediction.

Wong et al. [30] used econometric models to estimate gasoline consumption in Minnesota, USA, based on income, gasoline prices, vehicle market share, energy efficiency, and mileage. They concluded that income changes are key to predicting automotive energy demand. Badr et al. [31] investigated gasoline consumption in Lebanon, considering fuel prices and car registrations. They confirmed the statistical significance of gasoline prices on consumption but not on car registrations.

Jeon [32] used *General Circulation Models* (complex mathematical models used to simulate the behavior of the Earth's climate system) to estimate the impact of climate change on automotive fuel consumption in the USA. The author concluded that hot days increase gasoline consumption. However, there is no statistically significant effect on cold days, unlike what occurs in residential energy consumption.

Jaber et al. [33] and Al-Ghandoor et al. [34] carried out predictions based on linear regressions, where both studies use the estimated number of light vehicles, income level, and unit price of gasoline in their predictions. The first study found that without introducing diesel-powered cars, gasoline consumption would increase by 88.8% over a 10-year interval in Jordan and encourages this change. The second paper estimates that gasoline consumption in Jordan would increase at a rate of 1.81% per year.

Figueira et al. [35] used the SARIMA (Seasonal ARIMA) model to predict hydrous ethanol consumption based on Brazil's annual biofuel consumption series. For anhydrous ethanol, they employed a transfer function, an extension of ARIMA that incorporates exogenous variables, using gasoline consumption as the dependent variable and per capita GDP as the exogenous variable. They also evaluated the cross-correlation between GDP and gasoline consumption to select the best model. Dey et al. [36] similarly found ARIMA to be the most accurate for predicting annual gasoline demand in India.

Marquez et al. [22] presented univariate and multivariate approaches using LSTM to predict monthly ethanol consumption in Brazil. The univariate model used the historical hydrous ethanol consumption series, while the multivariate model included the following time series: i) ethanol consumption; ii) gasoline consumption; iii) flex-fuel vehicle fleet; iv) gasoline vehicle fleet; v) weighted average ethanol price; and vi) weighted average gasoline price. Although the multivariate LSTM model yielded the best results, their proposed prediction used the univariate model due to the difficulty of predicting the temporal behavior of the exogenous variables.

The most pertinent body of related works on the central theme of this paper is located in Subsection III-A, which focuses on predicting ethanol or other biofuel production. In contrast, Subsection III-B focuses on forecasting the consumption of gasoline or ethanol fuel, predominantly employing multivariate predictive models, with a few studies also integrating autoregressive models. Notably, both subsections present a significant gap in the literature concerning the application of DAI techniques to enhance the efficacy of AIM predictions.

IV. D-AI^2-M Methodology

The D-AI^2-M methodology is depicted in Fig. 2. This diagram outlines the step-by-step process for selecting, training,

and validating time series models tailored for monthly ethanol production prediction across the main Brazilian states and ethanol types. It details the data segregation, model training phases, hyperparameter optimization, and time series cross-validation techniques.

Step A in Fig. 2 represents the selection of ethanol production time series for inclusion in the experimental evaluation. This data selection considers production segregated by producing state and type of ethanol. The type of ethanol segregates the time series because the demand, process, and costs for the production of each product vary.

Step B describes the criteria for dividing each selected time series into subsets for training and cross-validation of the models. The initial training subsets excluded the most recent thirty-six months (last three years). Three time series cross-validation subsets were used, each cycle using the following 12 months for validation, with all previous data reassigned as the training set.

The proposed prediction process considers evaluating different autoregressive AIM subjected to different DAI preprocessing. This work refers to each model-preprocessing pair as D-AI^2-M. One of the objectives of the methodology is to measure the performance of D-AI^2-M by adopting a benchmarking process [37]. Step C1 of Fig. 2 adopts ARIMA as the reference model proposed by Salles et al. [37].

Steps C2, C2.1, and C2.2 cover the training of D-AI^2-M models. The autoregressive models presented in this article include ELM, MLP, SVR, Conv1D, and LSTM. These algorithms were trained and evaluated against the training base using the following preprocessing methods: DIFF, AN, and GMM. Step C2.2 details the hyperparameter optimization strategies, both general (applicable to all AIMs) and specific (tailored to each model type), used in the D-AI^2-M.

Step D presents the Rolling Forecast Origin methodology as the time series cross-validation criterion applied to ARIMA and D-AI^2-M models. This methodology involves creating a series of validation sets, where each test subset progressively moves forward in time. Correspondingly, each training subset includes all data before its respective test subset.

Step E of Fig. 2 consolidates the strategy for organizing and recording data related to the training and performance obtained by the trained models, including both benchmarking and D-AI^2-M models. Model performance was evaluated using the R^2 metric, which assumes higher values (close to one) for better-performing models and lower values (including negative ones) for poorer-performing models [38]. The performance of the benchmarking and D-AI^2-M models is documented for each training and testing subset, as defined in Step B. The performance of each model is determined considering the testing subset for each selected time series.

V. Experimental Protocol

This section presents the experimental evaluation according to the concepts and criteria covered in Section IV. This section is subdivided into two main topics: a) Dataset used, and b) Implementation.

A. Dataset

The data set used in the experimental evaluation comprised the time series of Brazilian production of anhydrous and hydrous ethanol, provided by ANP [6]. This time series has a monthly frequency, from January 2012 to December 2023, and includes the following attributes: i) Year, ii) Month, iii) Federation Unit, iv) Type of ethanol, and v) Production in m^3. To conduct the experiments, data were selected from the six Brazilian states with the highest ethanol production (SP, GO, MG, MT, MS, and PR) separated by type of produced ethanol.

B. Implementation

All experiments were conducted using the R programming language, mainly supported by the DAL Toolbox package [39]. The repository for this experimental evaluation can be accessed at https://github.com/cefet-rj-dal/DAI2M. This experimental evaluation used ARIMA models automatically adjusted by the Auto-ARIMA [40] function as benchmarking. The Auto-ARIMA is part of the Forecast package, developed in the R language [13].

For the D-AI^2-M training, hyperparameter optimization was conducted using the Grid Search methodology. This optimization process utilized common parameter ranges for all AIMs, including sliding windows and input size, as well as specific parameter ranges tailored to each type of AIM algorithm. These hyperparameter ranges are detailed in Table II. The table's notation ([i: j, k]) indicates a numerical sequence starting from i to j with a step size of k.

TABLE II
HYPERPARAMETERS FOR D-AI^2-M OPTIMIZATION

AIM	Hyperparameters
All 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

The criteria for creating the datasets used for time series cross-validation of the models are detailed in Fig. 3. The numbers in each rectangle represent the year of the respective data (12 months set). For each time series cross-validation cycle, light gray indicates that the rectangle is part of the training subset, and dark gray indicates that it is part of the validation subset.

VI. Results

This section presents the results from the ARIMA and D-AI^2-M models during cross-validation. The analysis aimed to: i) compare the average performance of ARIMA and each AIM for each time series after cross-validation (Table III); ii) compare the overall average performance of ARIMA with DAI (Table IV); iii) compare the local performance (for each series) of ARIMA and D-AI^2-M (Table V); and iv) identify the best global model, considering all selected series (Table VI). Starting from the second column of these tables, scenarios

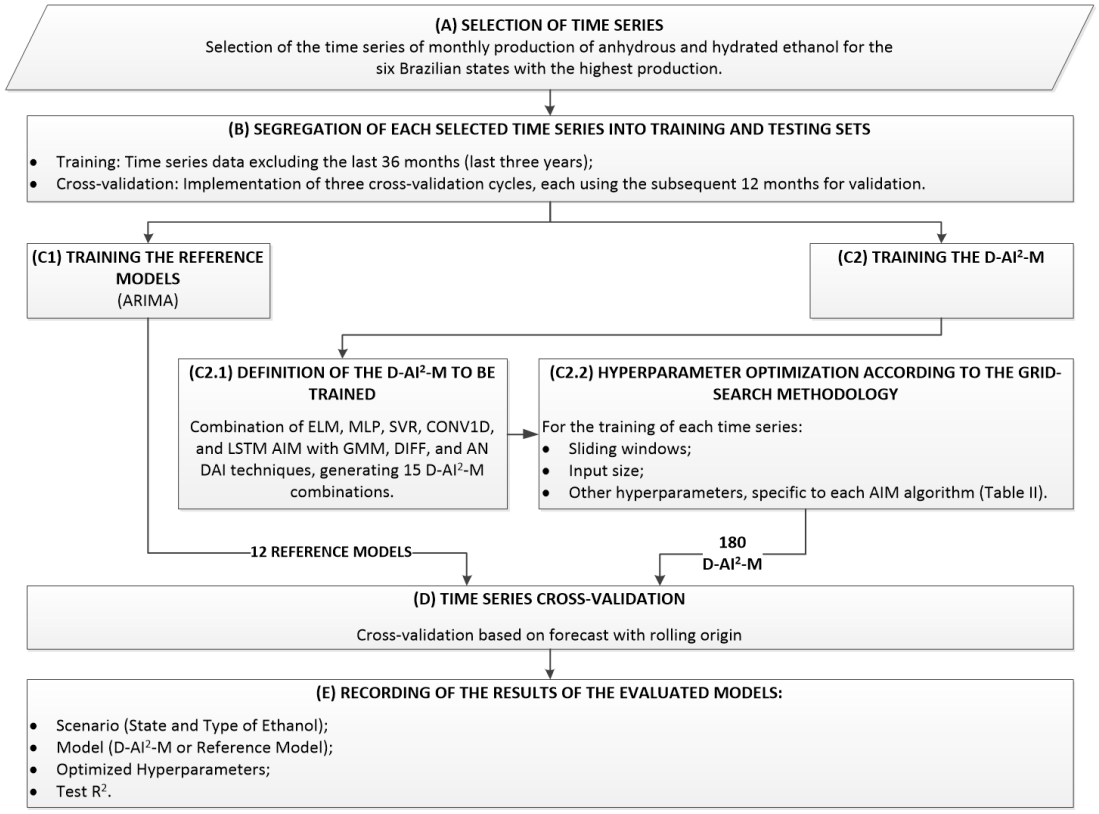

Fig. 2. D-AI^2-M Process - Selecting the best models for Brazilian monthly prediction of ethanol production.

Fig. 3. Details of time series cross-validation (rolling origin evaluation).

are presented, each identified by the state's abbreviation and ethanol type (*ANH* for anhydrous and *HYD* for hydrous). For instance, *GO ANH* represents anhydrous ethanol production in Goiás.

TABLE III
ARIMA AND AIM COMPARISON RESULTS (MEAN R^2 IN TIME SERIES CROSS-VALIDATION)

AIM / ARIMA	GO ANH	GO HYD	MG ANH	MG HYD	MS ANH	MS HYD
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
AIM / ARIMA	**MT ANH**	**MT HYD**	**PR ANH**	**PR HYD**	**SP ANH**	**SP HYD**
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

TABLE IV
ARIMA AND DAI COMPARISON RESULTS (MEAN R^2 IN TIME SERIES CROSS-VALIDATION)

DAI / ARIMA	GO ANH	GO HYD	MG ANH	MG HYD	MS ANH	MS HYD
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
DAI / ARIMA	**MT ANH**	**MT HYD**	**PR ANH**	**PR HYD**	**SP ANH**	**SP HYD**
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

In Table III, the results of the comparison between ARIMA and AIMs are presented. It can be seen that the ARIMA reference models perform better in most scenarios (seven out of twelve). For the scenario GO HYD, although ARIMA did not outperform, its performance is very close to that of the others. It is also worth noting that all models performed well for this time series. ARIMA did not perform best in the MS ANH, MT ANH, and MT HYD scenarios, but none of the models presented good results for these time series.

Table IV presents the results of the comparison between ARIMA and the average results of DAI preprocessing. Once again, it can be seen that the ARIMA reference models perform better in most scenarios (eight out of twelve). In addition to ARIMA, only DIFF emerged as the winner in some

TABLE V
ARIMA AND D-AI^2-M - LOCAL COMPARISON RESULTS (MEAN R^2 IN TIME SERIES CROSS-VALIDATION)

D-AI^2-M / ARIMA	GO ANH	GO HYD	MG ANH	MG HYD	MS ANH	MS HYD
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
D-AI^2-M / ARIMA	**MT ANH**	**MT HYD**	**PR ANH**	**PR HYD**	**SP ANH**	**SP HYD**
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

TABLE VI
ARIMA AND D-AI^2-M - GLOBAL COMPARISON RESULTS (MEAN R^2 IN EVALUATION AND TIME SERIES CROSS-VALIDATION)

D-AI^2-M / ARIMA	Selection: Evaluation $\pm$ SD (2021)	Cross-validation $\pm$ SD (2022 and 2023)
DIFF+ELM	**0.78 $\pm$0.14**	**0.60 $\pm$0.35**
DIFF+SVM	0.63 $\pm$0.24	0.50 $\pm$0.43
GMM+MLP	0.61 $\pm$0.37	0.30 $\pm$0.59
GMM+LSTM	0.61 $\pm$0.48	0.25 $\pm$0.60
AN+LSTM	0.54 $\pm$0.18	0.55 $\pm$0.46
ARIMA	0.52 $\pm$0.50	0.29 $\pm$0.87
GMM+ELM	0.50 $\pm$0.67	0.38 $\pm$0.62
AN+SVM	0.44 $\pm$0.28	0.36 $\pm$0.55
GMM+CONV1D	0.36 $\pm$0.69	-23.41 $\pm$110.78
AN+CONV1D	0.29 $\pm$0.56	-0.48 $\pm$3.02
DIFF+MLP	0.25 $\pm$0.80	0.51 $\pm$0.32
GMM+SVM	0.15 $\pm$1.91	0.44 $\pm$0.37
DIFF+CONV1D	0.07 $\pm$1.06	0.04 $\pm$0.93
AN+MLP	-0.37 $\pm$2.61	-37.1 $\pm$127.23
AN+ELM	-1.17 $\pm$3.63	-4.43 $\pm$15.99
DIFF+LSTM	-1.85 $\pm$7.23	-0.14 $\pm$2.95

scenarios (GO HYD, MS ANH, MT ANH, and PR HYD).

Table V presents the comparison results between the ARIMA models and the D-AI^2-M. In this comparison, the D-AI^2-M models outperform the ARIMA benchmark models in all evaluated scenarios. This analysis considers *local* performances specific to each time series [11]. The local results presented by the D-AI^2-M winners were superior to the best results identified in Tables III and IV. Only two ties existed between the D-AI^2-M and the ARIMA models for the ANI MG and ANI SP scenarios.

Finally, it is worth highlighting *DIFF+ELM*, which excelled in three scenarios during the local evaluation and achieved the best *global* performance in the experimental assessment [11]. This was measured by the R^2 metric using two approaches: i) *Selection*: with 2021 as the test base, and ii) *Validation*: cross-validation over 2022 and 2023. A paired Wilcoxon test [41] confirmed DIFF+ELM's superiority with a p-value of 6.387×10^{-5} compared to DIFF+SVM (second best global performance). An effect size evaluation [42] indicated a *high* difference between these models (r = 0.81).

An evaluation of residuals using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) [9] was conducted for both the ARIMA (benchmark) and DIFF+ELM (best global performance). Such ACF plots are shown in Fig. 4 (for hydrous ethanol production scenarios) and Fig. 5 (for anhydrous ethanol production scenarios). In such figures the horizontal dashed lines show the confidence intervals and the vertical lines the magnitudes of the respective lags. The evaluation considered the number of lags (excluding zero lags) that exceeded the confidence intervals for each model as a criterion. A smaller number of lags beyond the confidence interval indicates a model better fitted to the respective time series. The results of this analysis are recorded in Table VII, where it can be seen that DIFF+ELM showed superior performance in eleven of the twelve scenarios compared to ARIMA. As can be seen in this table, the PACF graphs presented similar behavior to the respective ACF graphs (in some cases varying proportionally the number of lags exceeding the confidence limits and their magnitudes). The PACF graphs were not shown in this article due to space constraints, but are available at https://github.com/cefet-rj-dal/DAI2M. Fig. 5 presents the only scenario (SP Anhydrous) where DIFF+ELM exhibited a greater number of lags exceeding the confidence intervals compared to the ARIMA model. However, it can be observed that, in general, ARIMA displayed larger magnitudes for the lags exceeding the confidence interval than DIFF+ELM, even in this scenario.

TABLE VII
ACF AND PACF ANALYSIS: LAGS BEYOND THE CONFIDENCE INTERVALS

SCENARIO	ARIMA ACF	DIFF+ELM ACF	ARIMA PACF	DIFF+ELM PACF
GO ANH	3	0	5	0
GO HYD	8	0	8	0
MG ANH	3	1	3	2
MG HYD	3	0	6	0
MS ANH	2	0	1	0
MS HYD	2	1	2	1
MT ANH	7	1	6	0
MT HYD	2	1	2	1
PR ANH	2	1	2	1
PR HYD	1	0	1	0
SP ANH	2	3	2	4
SP HYD	2	0	2	0

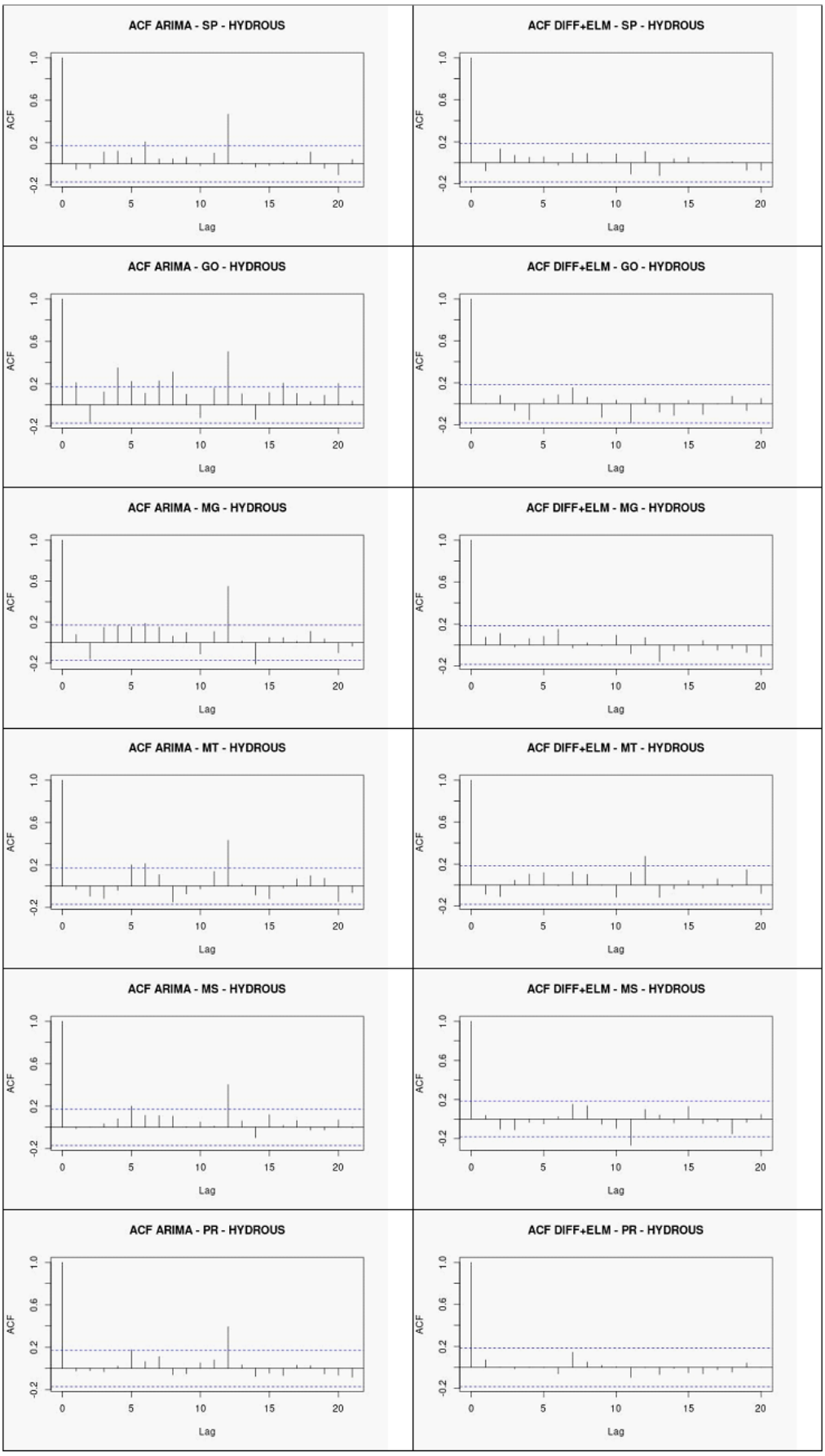

Fig. 4. ACF - ARIMA and DIFF+ELM residuals evaluation - Hydrous production scenarios.

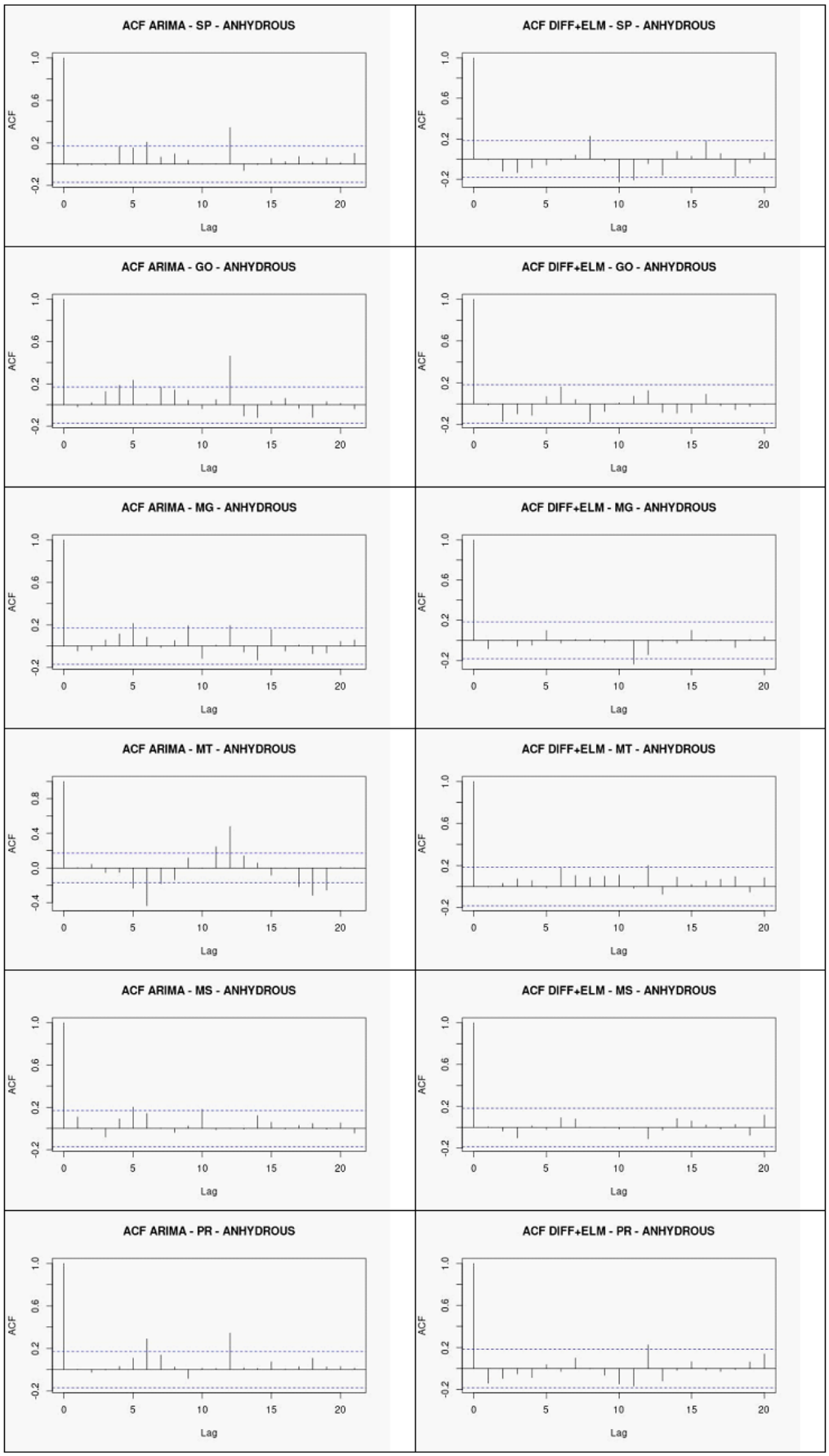

Fig. 5. ACF - ARIMA and DIFF+ELM residuals evaluation - Anhydrous production scenarios.

VII. Conclusion

This research shows that using DAI techniques was crucial in improving AIM performance in the evaluation scenarios. A key point is the importance of using time series cross-validation to analyze the models' robustness. This more robust analysis enhances the reliability of the results and the conclusions. However, this study did not aim to establish any causal relationships related to ethanol production.

The results encourage the use of AIM, especially when supported by an efficient model evaluation and selection methodology. Finally, D-AI^2-M effectively identified the best DAI and AIM combinations to select the most efficient models for each scenario (models with the best local performances). Additionally, the D-AI^2-M methodology identified the combination with the best global performance across all evaluated scenarios.

It is important to note that during part of the training period and the cross-validation, the world faced the COVID-19 pandemic. This event significantly affected consumption patterns and fuel production, including ethanol. Consequently, the time series reflected these impacts, influencing the models to varying degrees. Despite this challenge, D-AI^2-M efficiently predicted monthly ethanol production in Brazil.

Acknowledgments

The authors thank CNPq, CAPES, and FAPERJ for partially sponsoring this research.

References

[1] ANP, "Sales of petroleum derivatives and biofuels," https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos/vendas-de-derivados-de-petroleo-e-biocombustiveis, Tech. Rep., feb 2024.

[2] Renewable Fuels Association, "Annual Ethanol Production," https://ethanolrfa.org/markets-and-statistics/annual-ethanol-production, Tech. Rep., 2024.

[3] R. K. Niven, "Ethanol in gasoline: Environmental impacts and sustainability review article," *Renewable and Sustainable Energy Reviews*, vol. 9, no. 6, p. 535 – 555, 2005. doi: 10.1016/j.rser.2004.06.003

[4] E. Sadeghinezhad, S. N. Kazi, A. Badarudin, H. Togun, M. N. Zubir, C. S. Oon, and S. Gharehkhani, "Sustainability and environmental impact of ethanol as a biofuel," *Reviews in Chemical Engineering*, vol. 30, no. 1, p. 51 – 72, 2014. doi: 10.1515/revce-2013-0024

[5] A. L. da Silva and J. A. Castañeda-Ayarza, "Macro-environment analysis of the corn ethanol fuel development in Brazil," *Renewable and Sustainable Energy Reviews*, vol. 135, 2021. doi: 10.1016/j.rser.2020.110387

[6] ANP, "Production of biofuels," https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos/producao-de-biocombustiveis, Tech. Rep., apr 2024.

[7] S. G. Karp, J. D. C. Medina, L. A. J. Letti, A. L. Woiciechowski, J. C. de Carvalho, C. C. Schmitt, R. de Oliveira Penha, G. S. Kumlehn, and C. R. Soccol, "Bioeconomy and biofuels: the case of sugarcane ethanol in Brazil," *Biofuels, Bioproducts and Biorefining*, vol. 15, no. 3, pp. 899 – 912, 2021. doi: 10.1002/bbb.2195

[8] L. Gao, P. Lu, F. Qiao, J. Q. Li, Y. Zhang, and Y. Ren, "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*, vol. 6, 2022. doi: 10.1061/9780784484364.012 pp. 134 – 142.

[9] D. N. Gujarati, *Essentials of Econometrics*. SAGE, sep 2021. ISBN 978-1-07-185039-8

[10] M. H. Jarrahi, A. Memariani, and S. Guha, "The Principles of Data-Centric AI," *Communications of the ACM*, vol. 66, no. 8, pp. 84 – 92, 2023. doi: 10.1145/3571724

[11] P. Montero-Manso and R. J. Hyndman, "Principles and algorithms for forecasting groups of time series: Locality and globality," *International Journal of Forecasting*, vol. 37, no. 4, p. 1632 – 1653, 2021. doi: 10.1016/j.ijforecast.2021.03.004

[12] S. M. Al-Fattah, "A new artificial intelligence GANNATS model predicts gasoline demand of Saudi Arabia," *Journal of Petroleum Science and Engineering*, vol. 194, 2020. doi: 10.1016/j.petrol.2020.107528

[13] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, may 2018. ISBN 978-0-9875071-1-2

[14] E. Ogasawara, L. C. Martinez, D. De Oliveira, G. Zimbrão, G. L. Pappa, and M. Mattoso, "Adaptive Normalization: A novel data normalization approach for non-stationary time series," in *Proceedings of the International Joint Conference on Neural Networks*, 2010. doi: 10.1109/IJCNN.2010.5596746

[15] T. Tanaka, I. Nambu, Y. Maruyama, and Y. Wada, "Sliding-window normalization to improve the performance of machine-learning models for real-time motion prediction using electromyography," *Sensors*, vol. 22, no. 13, 2022. doi: 10.3390/s22135005

[16] E. Ogasawara, L. Murta, G. Zimbrão, and M. Mattoso, "Neural networks cartridges for data mining on time series," in *Proceedings of the International Joint Conference on Neural Networks*, 2009. doi: 10.1109/IJCNN.2009.5178615 pp. 2302 – 2309.

[17] G. Nasr, E. Badr, and C. Joun, "Backpropagation neural networks for modeling gasoline consumption," *Energy Conversion and Management*, vol. 44, no. 6, pp. 893 – 905, 2003. doi: 10.1016/S0196-8904(02)00087-0

[18] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, jul 2019. ISBN 978-0-13-461099-3

[19] T. Anh Tran, "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*, vol. 6, no. 4, pp. 317 – 339, 2021. doi: 10.1016/j.joes.2021.02.005

[20] J. Wang, S. Lu, S.-H. Wang, and Y.-D. Zhang, "A review on extreme learning machine," *Multimedia Tools and Applications*, vol. 81, no. 29, pp. 41 611–41 660, 2022. doi: 10.1007/s11042-021-11007-7

[21] B. Lim and S. Zohren, "Time-series forecasting with deep learning: A survey," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2194, 2021. doi: 10.1098/rsta.2020.0209

[22] J. P. Marquez, C. De Oliveira Ribeiro, E. R. Santoyo, and V. F. Fernandez, "Ethanol Fuel Demand Forecasting in Brazil Using a LSTM Recurrent Neural Network Approach," *IEEE Latin America Transactions*, vol. 19, no. 4, pp. 551 – 558, 2021. doi: 10.1109/TLA.2021.9448537

[23] S. Bhanja and A. Das, "Deep learning-based integrated stacked model for the stock market prediction," *Int. J. Eng. Adv. Technol*, vol. 9, no. 1, pp. 5167–5174, 2019. doi: 10.35940/ijeat.A1823.109119

[24] Z. Li, B. Zhou, and D. A. Hensher, "Forecasting automobile gasoline demand in Australia using machine learning-based regression," *Energy*, vol. 239, 2022. doi: 10.1016/j.energy.2021.122312

[25] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199 – 222, 2004. doi: 10.1023/B:STCO.0000035301.49549.88

[26] R. Fink and S. Medved, "Global perspectives on first generation liquid biofuel production," *Turkish Journal of Agriculture and Forestry*, vol. 35, no. 5, pp. 453 – 459, 2011. doi: 10.3906/tar-1005-905

[27] S. Badamchizadeh, A. J. Latibari, A. Tajdini, S. Pourmousa, and A. Lashgari, "Modeling Current and Future Role of Agricultural Waste in the Production of Bioethanol for Gasoline Vehicles," *BioResources*, vol. 16, no. 3, pp. 4798 – 4813, 2021. doi: 10.15376/biores.16.3.4798-4813

[28] L. Yu, S. Liang, R. Chen, and K. K. Lai, "Predicting monthly biofuel production using a hybrid ensemble forecasting methodology," *International Journal of Forecasting*, vol. 38, no. 1, pp. 3 – 20, 2022. doi: 10.1016/j.ijforecast.2019.08.014

[29] M. Melikoglu, "Demand forecast for road transportation fuels including gasoline, diesel, LPG, bioethanol and biodiesel for Turkey between 2013 and 2023," *Renewable Energy*, vol. 64, pp. 164 – 171, 2014. doi: 10.1016/j.renene.2013.11.009

[30] E. Wong, E. Venegas, and D. Antiporta, "Simulating the consumption of gasoline," *Simulation*, vol. 28, no. 5, pp. 145 – 152, 1977. doi: 10.1177/003754977702800505

[31] E. Badr, G. Nasr, and G. Dibeh, "Econometric modeling of gasoline consumption: A cointegration analysis," *Energy Sources, Part B: Economics, Planning and Policy*, vol. 3, no. 3, pp. 305 – 313, 2008. doi: 10.1080/15567240701232048

[32] H. Jeon, "The impact of climate change on passenger vehicle fuel

consumption: Evidence from U.S. panel data," *Energies*, vol. 12, no. 23, 2019. doi: 10.3390/en12234460
[33] J. O. Jaber, A. M. Al-Ghandoor, I. Al-Hinti, and S. A. Sawallha, "Prediction of energy consumption of passenger transportation and GHG emissions in Jordan," *International Journal of Global Warming*, vol. 4, no. 2, pp. 90 – 112, 2012. doi: 10.1504/IJGW.2012.048457
[34] A. Al-Ghandoor, J. Jaber, I. Al-Hinti, and Y. Abdallat, "Statistical assessment and analyses of the determinants of transportation sector gasoline demand in Jordan," *Transportation Research Part A: Policy and Practice*, vol. 50, pp. 129 – 138, 2013. doi: 10.1016/j.tra.2013.01.022
[35] S. R. Figueira, H. L. Burnquist, and M. R. P. Bacchi, "Forecasting fuel ethanol consumption in Brazil by time series models: 2006-2012," *Applied Economics*, vol. 42, no. 7, pp. 865 – 874, 2010. doi: 10.1080/00036840701720978
[36] B. Dey, B. Roy, S. Datta, and T. S. Ustun, "Forecasting ethanol demand in India to meet future blending targets: A comparison of ARIMA and various regression models," *Energy Reports*, vol. 9, pp. 411 – 418, 2023. doi: 10.1016/j.egyr.2022.11.038
[37] R. Salles, L. Assis, G. Guedes, E. Bezerra, F. Porto, and E. Ogasawara, "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*, vol. 2017-May, 2017. doi: 10.1109/IJCNN.2017.7966139 pp. 2338 – 2345.
[38] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, pp. 1 – 24, 2021. doi: 10.7717/PEERJ-CS.623
[39] E. Ogasawara, A. Castro, H. Borges, D. Carvalho, J. Santos, E. Bezerra, and R. Coutinho, "daltoolbox: Leveraging Experiment Lines to Data Analytics," jul 2023. [Online]. Available: https://cran.r-project.org/web/packages/daltoolbox/index.html
[40] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *Journal of Statistical Software*, vol. 27, no. 3, pp. 1 – 22, 2008. doi: 10.18637/jss.v027.i03
[41] F. Wilcoxon, "Probability tables for individual comparisons by ranking methods," *Biometrics*, vol. 3, no. 3, p. 119 – 122, 1947. doi: 10.2307/3001946
[42] M. Tomczak and E. Tomczak, "The need to report effect size estimates revisited. an overview of some recommended measures of effect size," *Trends in Sport Sciences*, vol. 1, no. 21, pp. 19–25, 2014.

Antonio Carlos Silva Mello is currently pursuing his master's degree in the Graduate Program in Production and Systems Engineering (PPPRO) at the Federal Center for Technological Education of Rio de Janeiro (CEFET/RJ). He holds a Business Analytics and Big Data postgraduate from the Getúlio Vargas Foundation (FGV) and a Mechatronics Engineering from the Federal University of Rio de Janeiro (UFRJ). His primary research interests lie in data science, particularly time series analysis.

Diego Carvalho (M'98-SM'19) was born in Rio de Janeiro, Brazil in 1970. He received his B.S. in Production Engineering from UFRJ and his M.Sc. and D.Sc. in Systems Engineering and Computer Science from PESC/COPPE. From 1993 to 1996, he was a Computer Research Assistant with the DELPHI Experiment at CERN. From 1997 to 2011, he was amongst the leading researchers of various EU-funded grid computing projects. Since 2006, he has been a professor at the Department of Production Engineering of CEFET/RJ, and his research interests include distributed systems, network engineering, parallel architectures, grid technologies, data mining, and big data. Dr. Carvalho is a member of the Brazilian Association of Production Engineering, the Brazilian Society for the Advancement of Science, and a senior member of IEEE.

Eduardo Ogasawara has been a professor at the Department of Computer Science at the Federal Center for Technological Education of Rio de Janeiro (CEFET/RJ) since 2010. He holds a D.Sc. in Systems and Computer Engineering from COPPE/UFRJ. Between 2000 and 2007, he worked in the Information Technology (IT) sector, gaining extensive experience in workflows and project management. With a strong background in Data Science, he is currently focused on Data Mining and Time Series Analysis. He is a member of IEEE, ACM, and SBC. Throughout his career, he has authored numerous published articles and led projects funded by agencies such as CNPq and FAPERJ. Currently, he heads the Data Analytics Lab (DAL) at CEFET/RJ, where he continues to advance research in Data Science.

Fabio Porto is a senior researcher at the National Laboratory of Scientific Computing, where he coordinates the Data Extreme Lab (DEXL). He holds a PhD and M.Sc. in Informatics from PUC-Rio and a Bachelor's degree in Mathematics and Informatics from the State University of Rio de Janeiro. After his PhD, he stayed as a Post-doc at the EPFL Database laboratory and was appointed Visiting Professor at the National University of Singapore from March to June 2020. His current research interests involve Databases and big Data Frameworks, integrating IA and Databases, and managing ML models and data.

Fernando Alexandrino holds a B.Sc. degree (CEFET/RJ, 2014) and a M.Sc. degree (COPPE/UFRJ, 2017) in Production Engineering. He is a professor at the Federal Institute of Technological Education of São Paulo (IFSP) and a Ph.D. student in the Postgraduate Program in Production Engineering and Systems at CEFET/RJ. His research interests involve Data Science, including pattern recognition and predictive modeling using weightless artificial neural networks.

Gustavo Guedes has been a professor at the Computer Science Department of the Federal Center for Technological Education of Rio de Janeiro (CEFET/RJ) since 2010. He earned his Doctorate (D.Sc.) in Computing and Systems Engineering from COPPE/UFRJ in 2015. His primary interests are in Affective Computing and Data Science, with significant experience in Affective Computing and Text Mining. Currently, he is leading the Affective Computing Lab at CEFET/RJ.

Jorge Soares holds a Ph.D. and a M. Sc. degree in Systems and Computer Engineering from COPPE/UFRJ, and a B.Sc. degree in Computer Science from UFRJ. He is a full professor at the Federal Center for Technological Education (CEFET/RJ). His main areas of interest are Data Science and Analytics, Database Systems, Data-Centric Artificial Intelligence (DCAI), and data integration. Recently, his fields of application have included agriculture, sports, and public health.

Lucas Giusti is currently pursuing a DSc. degree in Production and Systems Engineering at PPPRO/CEFET-RJ. He holds MSc. in Data Science (PPCIC/CEFET-RJ) and Exercise and Sports Sciences (PPCEE/UERJ). With 8 years of experience, he has worked as a Full and Senior Data Scientist on projects involving Soccer Performance and Talent Prediction, Predictive Maintenance, NLP, and Full Stack development. Lucas' research focuses on Concept Drift and Data Science applications in sports, particularly soccer and running.

Rafael Barbastefano holds a bachelor's degree in Production Engineering, a master's degree in Applied Mathematics and a doctorate in Engineering (Operations Research and Production Management) from the Federal University of Rio de Janeiro (UFRJ). He is a full professor at the Celso Suckow da Fonseca Federal Center for Technological Education and Scientific Director of the Brazilian Association of Production Engineering. He has experience in Social Networks, Operations Management, and Educational Technology, working on social network applications, technology prospecting, and distance education.

Tarsila Tavares holds a B.Sc. Statistics from ENCE. Her main areas of interest are Data Science and analytics and Data-Centric Artificial Intelligence (DCAI). She has 15 years of experience in the market, working on Data Science and Analytics projects and developing research in the area of DCAI in relevant companies in the financial sector, consultancies, and retail.