Agronomy Research **23**(2), 1074–1096, 2025
https://doi.org/10.15159/AR.25.066

Enhancing biogas production predictions using ARIMAX models on mixed silages

M. González-Palacio[1,*], L. González-Palacio[2], S. Villegas-Moncada[3], C. Arrieta-González[3], M. Luna-delRisco[3] and C. Arroyave-Quiceno[4]

[1]Universidad de Medellín, Faculty of Engineering, Department of Information Technology, Carrera 84 # 30-65, CO 050026 Medellín, Colombia
[2]Universidad EAFIT, Faculty of Engineering, Department of Product Design and Experience, Calle 49 # 7 Sur-50, CO 050022 Medellín, Colombia
[3]Universidad de Medellín, Faculty of Engineering, Department of Energy, Carrera 84 # 30-65, 050026 Medellín, Colombia
[4]Universidad de Medellín, Faculty of Engineering, Department of Environmental Sciences, Carrera 84 # 30-65, CO 050026 Medellín, Colombia
[*]Correspondence: magonzalez@udemedellin.edu.co

Received: February 15th, 2025; Accepted: June 19th, 2025; Published: July 14th, 2025

Abstract. Biogas production as a renewable energy source is gaining more attention from different actors in the energy sector due to the use of different residual products for its generation. This interest also comes from the agricultural sector. A typical crop used for biogas production is maize, which poses environmental challenges related to soil erosion and nutrient depletion. Furthermore, land use changes can also reduce biodiversity and attract pests. An increasing number of strategies to diminish these issues rely on combining maize with other leguminous plants, improving the nutritional silage profiles, and potentially enhancing biogas production. Nonetheless, adopting these new approaches remains limited since the farmers hesitate to invest in new technologies without clear and quantifiable improvements. In this regard, in this study, we propose time-series-based models to predict biogas and methane production based on the silage features of crops and the time-series data. In particular, we fitted models based on Autoregressive Integrated Moving Average with eXogenous variables (ARIMAX) to capture the temporal dependencies, aiming to characterize the methane volume and methane concentration accurately. We used a previously validated measurement campaign, which included other anaerobic digestion variables like volatile solids, crude protein, cellulose, and hemicellulose, among others, from crops of maize and mixed maize-legume silages, along with the production of biogas and methane, with a sample period in days. The reactor was a 5 L fermenter operated at 40 °C with manual mixing daily. It used inoculum and silage, with a 21-day delay before measurement. Biogas volume was recorded using a measuring cylinder, and composition was analyzed with a Dräger X-am 8000. We tested our ARIMAX-based models regarding their goodness of fit using the determination coefficient R^2 and the Root Mean Square Error (RMSE). In the case of the methane volume, we obtained an R^2 of 0.92 and an RMSE of 0.001 liters, and for the case of methane concentration, our models exhibited an R^2 of 0.908 and an RMSE of 0.85%. Our promising models help farmers, researchers, and policymakers to accurately

characterize and forecast biogas and methane production as promising renewable energy generation technologies.

Key words: anaerobic digestion, biogas production, time series, maize-legume silage, time-series forecasting.

INTRODUCTION

The global demand for energy sources has increased dramatically regarding the growing of production industries, data centers, computing for processing Artificial Intelligence (AI) based solutions, among others (Tiismus et al., 2025). In this regard, different alternative and renewable energy strategies have arisen to diversify the energy matrix, aiming to diminish the dependence on classical generation techniques based on fossil fuels, nuclear, or hydroelectric power plants (Mamidala et al., 2023). Among these approaches, anaerobic digestion has emerged as a suitable and promising option for biogas production to be further employed for thermal energy generation (Liu et al., 2021). This biological process is based on the use of organic materials that are decomposed by microorganisms in an oxygen-free environment, resulting in methane, carbon dioxide, and minimal traces of other gases depending on the chosen feedstocks (Mao et al., 2015). Furthermore, anaerobic digestion offers unique advantages, like the types of feedstocks that can be used to produce biogas, including livestock manure, food wastes, and sewage sludge (Thakur et al., 2023). For instance, pig manure is commonly used since its broad production in dedicated farms, making this feedstock highly available and biodegradable (Ferreira et al., 2024). Nonetheless, animal-based feedstocks can be very variable according to the characteristics of the surrounding environment, as in other processes (González-Palacio et al., 2024). The factors include climate, animal diet, housing systems, and management practices. Recent studies have emphasized that these environmental conditions significantly influence the physicochemical composition of animal-derived materials, such as manure, fat, and by-products, which in turn affect their suitability for bioenergy production or nutrient recovery (Okolie et al., 2023). This variability challenges the prediction of methane production and concentration.

In that way, promising alternatives of feedstocks based on crops started capturing the attention of the productive sector since they offer several advantages over other substrates, like the high level of biomass yield and convenient biochemical composition, among others (Hoang et al., 2022), which refers to the high content of fermentable sugars, cellulose, and hemicellulose, which are ideal for biofuel production due to their ease of conversion and high energy potential (Himmel et al., 2007). In particular, maize offers the advantage of high predictability since it can be produced to be consistent, facilitating a stable Anaerobic Digestion (AD) process. The consistency of maize silage is characterized by a relatively uniform chemical composition, particularly in terms of its carbohydrate content, which includes cellulose and hemicellulose. This uniformity contributes to predictable and stable biogas production rates and yields and stable AD performance (Kaparaju & Rintala, 2005). However, the cultivation of maize poses different environmental challenges, such as soil erosion and nutrient depletion (Ologunde et al., 2025). These issues can be mitigated by intercropping maize with

leguminous plants, enhancing the soil structure, fertility, and nutritional profile, improving the biogas production (Kintl et al., 2024). Nonetheless, the prediction of biogas production to adequately forecast the performance under different configurations is a time-consuming and expensive process that can be threatened by multiple sources, such as weather variations or pests (Pfordt & Paulus, 2025).

For this reason, active research is being conducted to develop reliable methods for predicting biogas production in nonlinear and highly variable AD processes, with the goal of identifying the best operational parameters, such as temperature, retention time, and feedstock mixture ratios, to enhance system efficiency and methane yield. Previous attempts at digester performance prediction employ mostly machine-learning models whose operation may not be traced back directly from the physical processes of the digester. The authors in (Kim et al., 2025) depicted the everyday evolution of a whey-fed lab reactor on a graph-convolutional network. The study used laboratory-scale anaerobic digesters with a 50 mL working volume, operated at 37 °C, and fed with whey-based substrate every three days to maintain a hydraulic retention time of 30 days. Different organic loading rates were applied, including high, medium, and low levels, with fluctuating patterns such as stepwise and abrupt changes to mimic real-world operational variations. Although the network simulated gas-rate dynamics highly accurately (MSE = 0.01) and identified shifts in the microbial community abundance with an R^2 of 0.72 and a microbial community composition with an R^2 of 0.87, the large number of latent parameters prevents a straightforward explanation of such associations.

The study conducted in (Zou et al., 2024) proposed a machine learning-based prediction model of biogas production in full-scale dry anaerobic digestion of kitchen food waste. The study employed operational data collected for 1.5 years from four full-scale digesters, including parameters such as biomass amount, addition of centrifuged sludge, liquid level, digester capacity, amount of product, and biochemical parameters such as pH. Eight machine learning algorithms were evaluated, of which CatBoost was the most accurate with R^2 values from 0.604 to 0.915 for biogas prediction. Although the high predictive performance, the complexity of the model, and the many latent parameters complicate a direct understanding of the inter-variable relationships.

The research carried out in (Avinash & Mishra, 2024) investigated the predictive potential of AI-based and kinetic models for AD biogas production. Batch digestion trials under various moisture content, organic load, and operational parameters like temperature, pH, and retention time were carried out to train the models. Different models, like artificial neural network (ANN), adaptive neuro-fuzzy inference systems (ANFIS), were trained and tested based on experimental biogas yield data. ANFIS network produced biogas yield with high precision (RMSE = 0.670, R^2 = 0.999) compared to traditional kinetic models that also provided good fits. Although the models were able to identify general trends, the dense input variable and biogas output relation within particular AI models had numerous parameters and hence may be hard to interpret as direct underlying biological processes.

The investigation reported in (Farzin et al., 2024) evaluated the efficacy of auto-tuning machine learning models for biogas yield prediction from a municipal wastewater treatment plant on a full scale. Data obtained from operation experiments collected from South-Tehran municipal wastewater treatment plant were utilized for parameters such as

influent and effluent volatile solids, pH, temperature, and organic loading rates. Different combinations of hyperparameter tuning and feature selection were performed using genetic algorithms and particle swarm optimization for enhancing model performance. Support vector regression (SVR) and ANNs were used with the models. The most precise model was formed using SVR with $R^2 = 0.77$ for test and 0.49 for RMSE. Despite this achievement, the complex relationship among the selected parameters discourages simple interpretability of the relationships among operating variables and biogas production.

The approach in (Geng et al., 2024) proposed a prediction model of food waste anaerobic digestion biogas production using an augmented mix-up data augmentation strategy and an improved global attention Long-Short Term Memory (LSTM) network. The data were derived from two AD food waste reactors with actual production data, including parameters such as pH, volatile fatty acids, total solids, volatile suspended solids, ammonia nitrogen, and chemical oxygen demand from a six-month monitoring period. Different feeding regimes and organic loading conditions were mimicked to simulate operational variances. Artificial training samples were generated by the proposed model via mix-up in an attempt to counteract the data sparsity issue, and utilized an improved global attention scheme in the LSTM architecture for better modeling of temporal dependency. Performance in prediction was adequate ($R = 0.988$), and the model strongly predicted the actual daily biogas production (MSE = 0.002) and was capable of providing recommendations for feed adjustment to optimal energy output. However, the interpretability of each input feature is lost with added complexity due to the attention mechanism and data augmentation methods.

While these studies show that data-driven models are capable of representing highly variable processes, they are subjected to a key limitation since they are based on non-interpretable architectures, so practitioners lack explicit information about how different predictor variables can affect the prediction. Few recent studies have revisited statistical time-series models. The authors in (Prasad et al., 2023) compared SARIMAX and ARIMA with a variety of machine-learning baselines on 117 days of batch data. The results showed that plain ARIMA had the lowest error of prediction (RMSE = 3.26 L day^{-1}), indicating that plain parametric models perform similarly to black-box learners when faced with common temporal autocorrelations.

The summary of the reviewed approaches is shown in Table 1. where the following can be noticed. While these efforts demonstrate that data-driven models are able to track highly variable processes, they both possess two limitations that are of particular importance for crop-based digestion: (i) none of them deals with dedicated energy crops or mixtures of forage, though these substrates form a growing fraction of agricultural biogas plants, and (ii) all are based on non-interpretable architectures, thus practitioners have no explicit guidance on how fibre, starch or protein content affects the prediction. In this paper, we present a method based on Autoregressive Integrated Moving Average with eXogenous variables (ARIMAX) models to estimate methane volume and content for several silage profiles. We used an extensive measurement campaign in (Kintl et al., 2024) where different silages were prepared for the experiment, including the monocultural

maize silage (MA) and the mixtures of maize with white sweet clover (MA+WSC), white lupin (MA+LU), and fodder vetch (MA+VE) in a wet weight ratio of 70:30.

Table 1. Comparison of techniques used to forecast biogas production in anaerobic digesters

Author	Variables considered to forecast biogas production	Techniques	Black-box?
(Kim et al., 2025)	Temperature (37 °C), hydraulic retention time (30 d), organic loading rate (high/medium/low, stepwise & abrupt changes), feeding frequency (every 3 d), microbial community composition	Graph-Convolutional Network (GCN)	Yes
(Zou et al., 2024)	Biomass amount, addition of centrifuged sludge, liquid level, digester capacity, amount of product removed, pH	CatBoost gradient-boosted decision trees	Yes
(Avinash & Mishra, 2024)	Moisture content, organic load, temperature, pH, retention time	ANFIS / ANN	Yes
(Farzin et al., 2024)	Influent & effluent volatile solids, pH, temperature, organic loading rate	Support Vector Regression (SVR) + GA/PSO tuning	Yes
(Geng et al., 2024)	pH, volatile fatty acids, total solids, volatile suspended solids, ammonia nitrogen, chemical oxygen demand	Global-attention LSTM with mix-up augmentation	Yes

Parameters of measurement considered during the experiments involved the silage chemical composition including measured variables like volatile solids, neutral detergent fiber, acid detergent fiber, crude fiber, starch, cellulose, hemicellulose, crude protein, lipids, and acid detergent lignin (Palacio et al., 2017); while volume and composition of the produced biogas were analyzed for methane yield through fermentation tests. Other physical parameters measured were the density of the silage and some other operational variables, namely, temperature of incubation, period of fermentation, and amount of silage used (Gonzalez-Palacio et al., 2018). Our original contributions are three-fold:

1. A feature extraction process where we analyze how different variables affect the production and concentration of methane using statistical indexes.
2. A modeling stage where we fit a parametric ARIMAX that considers the predictor variables that highly influence the AD process and methane production.
3. We interpret how the different variables affect the process of methane production, providing valuable insights on how to improve anaerobic digestion processes.

The rest of this paper is organized as follows: Section II describes the database we used to determine each variable's statistical features. Section III is divided into two subsections. The first subsection explains the feature analysis and extraction, and the second subsection shows the proposed models. Section IV shows the results after applying the proposed methodology and analyzes their practical implications. Finally, Section V concludes the paper.

EXPERIMENT & DATABASE DESCRIPTION

The experiment in (Kintl et al., 2024) consisted of assessing biogas production from different silage mixtures, namely maize and legumes. The silages were made in mini-silos, with the silage mixtures being monocultural maize silage (designated as MA) and mixtures of maize with white sweet clover (MA+WSC), white lupin (MA+LU), and fodder vetch (MA+VE) in a fresh weight ratio of 70:30. The maize for the silages was cut manually at a stubble height of 18 cm, shredded into pieces of approximately 15–20 mm using a cutter (Deutz-Fahr MH 6505), then mixed with a bacterial inoculant (Silo Solve EF by Chr. Hansen Holding Ltd.) and filled into mini-silos and fermented. After that, the authors subjected the silages to 90 days of incubation in anaerobic conditions. Then, the mini-silos were sampled for subsequent analyses. For the fermentation, 24 batch fermenters of five liters each were used, preserving the anaerobic conditions throughout the experiment. There were three fermentation systems with eight fermenters each, totaling 24 fermenters. In each system, two fermenters were controls, and the remaining six fermenters were dosed with silage samples. Since there were four silage variants, each was tested in three replicate fermenters across all systems. After the incubation period of 90 days, the mini-silos were opened, samples were homogenized, frozen, and transported for chemical analyses and fermentation tests. Regarding sampling frequency, the samples were collected once after the 90-day incubation.

The setup permitted daily measurement of biogas production, with the produced biogas displacing a salt-saturated solution from the measuring cylinder into an expansion tank. The biogas generated was passed through an instrument (Dräger X-am 8000) for its composition, especially methane content. The fermenter conditions were carefully controlled, where the temperature was at 40 °C ± 0.2 °C provided by water baths. From the fermentation process, the volume and composition of the produced biogas were determined. Besides, the efficiency and yield of the biogas and methane production were also assessed from the varying silage mixtures. The variables measured in the experiment trials are as follows:

1. **SUBSTRATE:** The type of silage being analyzed (e.g., maize, mixed silage with legumes).
2. **DAY:** The day of sampling during fermentation.
3. **DM:** Dry Matter (%), representing the portion of the silage that remains after water is removed.
4. **VS:** Volatile Solids (%), a measure of the organic matter in the silage that can generate biogas.
5. **NDF:** Neutral Detergent Fiber (%), indicating the structural carbohydrates present in the plant material.
6. **ADF:** Acid Detergent Fiber (%), a measure of the resistant fibrous fraction containing cellulose and lignin.
7. **CF:** Crude Fiber (%), representing the indigestible portion of the feedstuff.
8. **STARCH:** Starch (%), the amount of starch present in the silage, which is a key fermentation substrate.

9. **ASH:** Ash (%), representing the inorganic mineral content remaining after combustion of the organic matter.

10. **CELLULOSIS:** Cellulose (%), the content of cellulose, a major component of the plant cell wall.

11. **HEMI-CELLULOSIS:** Hemicellulose (%), the content of hemicellulose, another digestible structural carbohydrate.

12. **CP:** Crude Protein (%), measuring total protein content available in the silage.

13. **LIPIDS:** Lipids (%), representing the fat content present in the silage.

14. **ADL:** Acid Detergent Lignin (%), measuring the lignin content, which affects digestibility.

15. **BIOGAS:** Biogas Volume (L), the total volume of gas produced during anaerobic digestion.

16. **METHANE:** Methane Volume (L), indicating the volume of methane gas produced as part of the biogas.

17. **METHANE_CONTENT:** Methane Content (%), the concentration of methane in the biogas, highlighting the fermentation efficiency.

Furthermore, Table 2 shows the descriptive statistics regarding the mean, standard deviation, minimum, maximum, and quartiles. These statistics were calculated using the entire dataset, encompassing the maize and the diverse silage mixtures, illustrating the distributions and variability of the key chemical parameters across the experiments.

Table 2. Descriptive statistics of the AD process

	Mean	Std. Dev	Min	Q1	Median	Q3	Max
DAY	11	6.092	1	6	11	16	21
DM	32.222	1.136	30.53	31.783	32.33	32.77	33.7
VS	95.65	0.519	94.78	95.507	95.885	96.028	96.05
NDF	33.825	3.347	30.17	32.03	32.95	34.745	39.23
ADF	22.807	2.332	20.41	21.438	22.1	23.47	26.62
CF	17.44	2.812	15.19	15.73	16.175	17.885	22.22
STARCH	24.782	7.8	15.57	18.322	24.285	30.745	34.99
ASH	4.35	0.519	3.95	3.973	4.115	4.492	5.22
CELLULOSIS	22.727	3.342	19.49	19.97	21.775	24.533	27.87
HEMI-CELLULOSIS	12.375	0.826	11.04	12.037	12.66	12.997	13.14
CP	11.28	1.719	9.03	10.012	11.325	12.593	13.44
LIPIDS	2.81	0.404	2.25	2.625	2.805	2.99	3.38
ADL	0.568	0.268	0.32	0.35	0.48	0.698	0.99
BIOGAS	**0.652**	**0.18**	**0.145**	**0.605**	**0.658**	**0.713**	**1.026**
METHANE	**0.254**	**0.085**	**0.022**	**0.22**	**0.254**	**0.304**	**0.383**
METHANE_CONTENT	**38.815**	**10.441**	**12.22**	**28.689**	**37.537**	**42.601**	**55.106**

This statistical overview can be used to understand and model their influences on biogas production. From this table, the following can be noticed. The DM content did not exhibit high variability, with an average of 32.22% and a standard deviation of 1.136, which denotes homogeneity in the composition of the substrate. Similarly, the VS, which is important for methane production, had little variation, with an average of 95.65% and a small range between 94.78% and 96.05%, which also suggests homogeneity in the organic matter present and favorable to microbial degradation. On the other hand, NDF

and ADF, indicative of carbohydrate substrate structure, were highly varied and ranged between 30.17% and 39.23% for NDF and 20.41% and 26.62% for ADF, thus showing variation in the composition of materials subjected to experiments, especially with respect to fiber contents. The same behavior was exhibited by the CF contents, which varied in the range of 16.175%, with the highest value as high as 22.22%, a fact suggesting variation in digestibility for samples. Besides, starch contents, impacting the immediate availability of fermentable carbohydrates, were very variable, with a mean of 24.78% and a standard deviation of 7.8%, ranging from 15.57% to 34.99%, suggesting variation in silage composition, which might impact fermentation processes. Furthermore, ash contents were also variable, highly uniform, and showed values that cluster around 4.35%, reflecting the incorporation of inorganic matter. The cellulose and hemicellulose contents, contributing to the slow biomass decomposition, were relatively variable in their content. Cellulose ranged between 19.49 and 27.87%, with an average of 22.727%, while hemicellulose ranged between 11.04 and 13.14%, indicating a relatively uniform composition in this portion of the fiber. The crude protein content, which is important for microbial processes, averaged 11.28 and ranged between 9.03 and 13.44%, indicating variable substrate compositions. Even though present in lesser concentrations, the range for lipids extended between 2.25 and 3.38%, with an average of 2.81%, and could impact methane yield through high-energy contents. The lignin of acid detergent, which represents the resistant biomass, was characterized by minimum values, with an average of 0.568 and a high of 0.99%. This indicates that lignin, which is one of the limiting substrates in anaerobic digestion, is at a minimum concentration and might, therefore, encourage higher biodegradability.The yield range for biogas production was from 0.145 to 1.026 liters (l), averaging (l). Similarly, the methane yield ranged between an average of 0.254 L and a high of 0.383 L. Methane in biogas had a high variance between 12.22% and 55.1%, averaging at 38.8%. This suggests that variable substrate compositions impact methanation concentrations notably. Together, the magnitude and spread of biogas, methane, and methane content (bolded in Table 2) provide a realistic, internally consistent dataset that reflects both the transient and the quasi-steady phases of AD, making it well suited for performance analysis and for testing predictive models that must cope with highly dynamic digester conditions.

METHODOLOGY

The modeling process is depicted in Fig. 1 and comprises four stages. The first stage is the Extraction, Transformation, and Load (ETL) step to transform the variables into a set of suitable measurements that can be used for modelling purposes. The second stage analyzes which predictor variables effectively explain the variability of the objective variables and selects the subset of them that is useful for the subsequent models. The third stage deals with the modeling, where ARIMAX-based models are fitted to predict methane volume and methane content. Our modeling framework is built upon the ARIMAX(p,d,q) model, an advanced version of the standard ARIMA model incorporating external predictors. In this context, p refers to the autoregressive terms representing the effect on the target variable at present due to values from the past, d is the number of differentiations necessary to fulfill the stationarity assumption, and q refers to the count of moving-average terms describing serially correlated shocks. The ‘X’ in ARIMAX indicates that, in addition to its own history and past of error

forecasts, the series is regressed on exogenous covariates (in this case, silage composition measurement) such that the model is able both to capitalize on internal temporal relationships and quantify the direct impact of external predictors on biogas and methane output. The parameters are estimated simultaneously (typically by maximum likelihood), and the selection of *p, d*, and *q* is based on standard diagnostic statistics such as the autocorrelation and partial autocorrelation functions. Finally, the fourth stage uses the models' predictions to analyze the results and practical implications. The following subsections detail each stage.

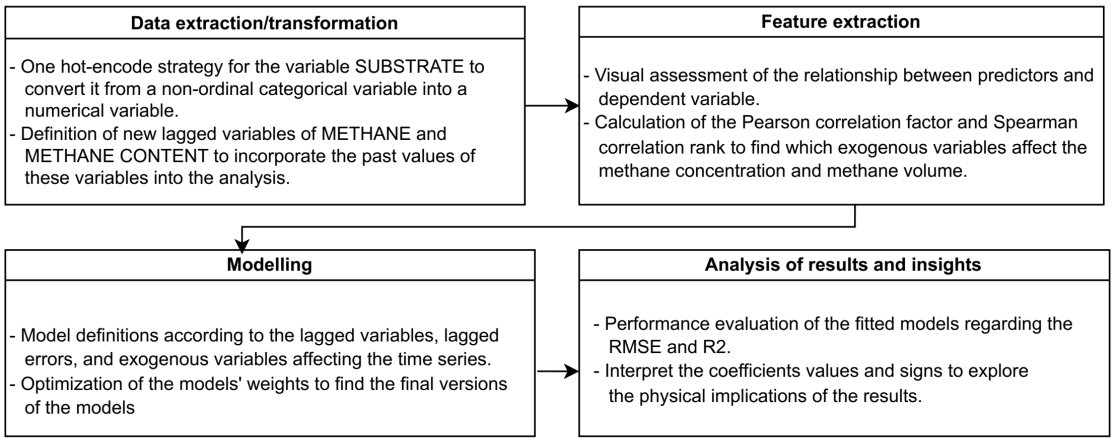

Figure 1. Modeling process that consists of four stages: data extraction and transformation, feature extraction, modeling, and analysis of results.

Data extraction and transformation

The first step we carried out to prepare the database includes two tasks: *i)* a one-hot encoding strategy to convert a non-ordinal categorical variable into a numerical one, like in the case of SUBSTRATE, whose order and values do not exhibit a particular order, and *ii)* a lagging strategy for the variables METHANE and METHANE CONTENT since past values of these variables affect the current and future state of them.

One-Hot encoding

This technique is used to convert categorical variables into numerical representations without inserting a false ordinal relation between categories. If a categorical variable X has k categories, we can transform it into a binary matrix representation where each category represents a binary feature. Mathematically, the variable X can be represented as

$$X = \{\mathrm{MA}, \mathrm{MA} + \mathrm{LU}, \mathrm{MA} + \mathrm{VE}, \mathrm{MA} + \mathrm{WSC}\}, \quad (1)$$

where each element in the set represents a particular combination of the maize and a certain leguminous. We can create a one-hot encoding representation using an indicator function by

$$X_i^{(j)} = \begin{cases} 1, \text{if the sample } i \text{ belongs to category } j \\ 0, \text{otherwise} \end{cases} \quad (2)$$

where $X_i^{(j)}$ represents the presence (1) or absence (0) of category j in sample i. This transformation is required since the considered models cannot process or analyze

categorical variables directly. It avoids the algorithms assuming a hierarchical or categorical relationship among the categories, which is incorrect in this context.

Lagged variables

Since the methane concentration and methane volume can be modeled using time-series analyses, the incorporation of lagged variables is crucial to enhance the predictability of the fitted models. This dependence can be mathematically modeled as follows. Let Y_t be a time-series variable (e.g., METHANE or METHANE_CONTENT) where t represents time (days in this case). A lagged variable is

$$Y_{t-k} = \text{Value of } Y \text{ at time } t-k, \tag{3}$$

where k represents the lag order (e.g., 1, 2, ..., n days before). In that way, each lag indicates a new variable that can be considered for the time-series analysis for feature extraction purposes.

Features extraction

The second step deals with the relationship between predictor variables with respect to the dependent variable. To this end, we conduct the following steps. First, we plot each independent variable versus the METHANE and METHANE_CONTENT to understand if a predictable pattern can be found from these relationships. After defining these regions, we use the Pearson correlation factor which describes the linearity between the variables within the range of [−1, 1]. That is, *i)* a value close to +1 means a strong positive correlation; *ii)* a value close to -1 means a strong negative correlation; and *iii)* a value close to zero means an insignificant correlation.

Besides this, the Spearman correlation coefficient measures whether there is a monotonic nonlinear function that captures the patterns and relationships between two variables. It is performed by ranking data from lowest to highest, assigning sequential ranks across every row in the dataset, calculating a difference d for both the predictor and dependent variables at every rank, and computing the rank correlation coefficient ρ. Our design criterion was to consider the variables whose Pearson and Spearman indices were above 0.5, meaning that the chosen variables can explain more than 50% of the variability of the corresponding dependent variables.

On the other hand, we also have to analyze how the past values of METHANE and METHANE_CONTENT can affect the current values of both variables, that is, how the lagged versions of them can be used to forecast their current and future values. It can be performed using the following analysis: *i)* first, we determine if the time series for both variables are stationary, using the Augmented Dickey-Fuller (ADF) test, *ii)* in case of non-stationarity, differentiate the time series until stationarity, *iii)* obtain the Partial Autocorrelation Function (PACF) and the Autocorrelation Function (ACF), and *iii)* with these results, determine the lagged variables and past errors that affect the current and future predictions. Regarding the ADF test, we can calculate the following expression for p lags by

$$\Delta Y_t = \delta Y_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta Y_{t-1} + \epsilon_t, \tag{4}$$

where $\Delta Y_t = (\rho - 1) Y_{t-1}$ represents the first-order difference, δ is a coefficient obtained using Ordinary Least Squares which represents how the prior value Y_{t-1} affects the

current value Y_t, α_i are the coefficient for p lagged values of Y_t, and ϵ_t is zero-mean independent and identically distributed gaussian noise. Then, we calculate the ADF statistic by

$$\text{ADF statistic} = \frac{\rho}{\text{SE}(\rho)}, \tag{5}$$

where SE(ρ) is the standard error of ρ. The ADF statistic is then compared with the critical value of 0.05 (95% confidence) and if it is less than the critical value, we conclude that the time series is stationary. In case that the series is not stationary, we compute the first difference as

$$Y_{\text{lag1}} = Y_i - Y_{i-1}, \tag{6}$$

where Y_i is the i^{th} value of the series and Y_{i-1} is the prior value of the series. The same procedure can be implemented for more lags if needed.

Modelling

Once we have selected the features that are essential in the prediction of METHANE and METHANE_CONTENT, we fit a parametric method (ARIMAX). The ARIMAX(*p, d, q*) model is a time-series model that aims to incorporate the dependence of the dependent variable on their p past values, q past forecasted errors, d differentiations over the dependent variable, and the influence of the exogenous variables. It can be represented by

$$Y_t^* = \beta_0 + \sum_{k=1}^{K} \beta_k X_{t,k} + \sum_{i=1}^{p} \phi_i Y_{t-i}^* + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t, \tag{7}$$

where Y_t^* represents the differenced series after applying d levels of differencing to achieve stationarity. The term β_0 is the intercept, β_k represents the coefficients associated with the exogenous variable $X_{t,k}$, considering that there are K exogenous variables. The coefficients ϕ_i correspond to the autoregressive terms that capture the dependence of the series on the past values. The moving average terms are weighted using θ_j, accounting for the past forecast errors on the current observation, and ϵ_t is the error noise. We use the PACF and ACF to determine the values for p and q. We use the Pearson/Spearman indexes to determine the exogenous variables that affect METHANE and METHANE_CONTENT from Table 2. Finally, we differentiate the series until the ADF test indicates stationarity.

RESULTS

This section presents the results of the proposed framework for modeling METHANE and METHANE_CONTENT. We used Anaconda Python for the statistical analysis regarding correlations, stationarity, and ARIMAX modeling. The time series are depicted in Fig. 2. From these trends, it can be noticed that the production and concentration of methane exhibit two stages. First, the variables rapidly increase, and then they tend to stabilize. The changes in the substrates influence the system's capacity to produce methane. In that way, we will provide ARIMAX-based models for both METHANE and METHANE_CONTENT.

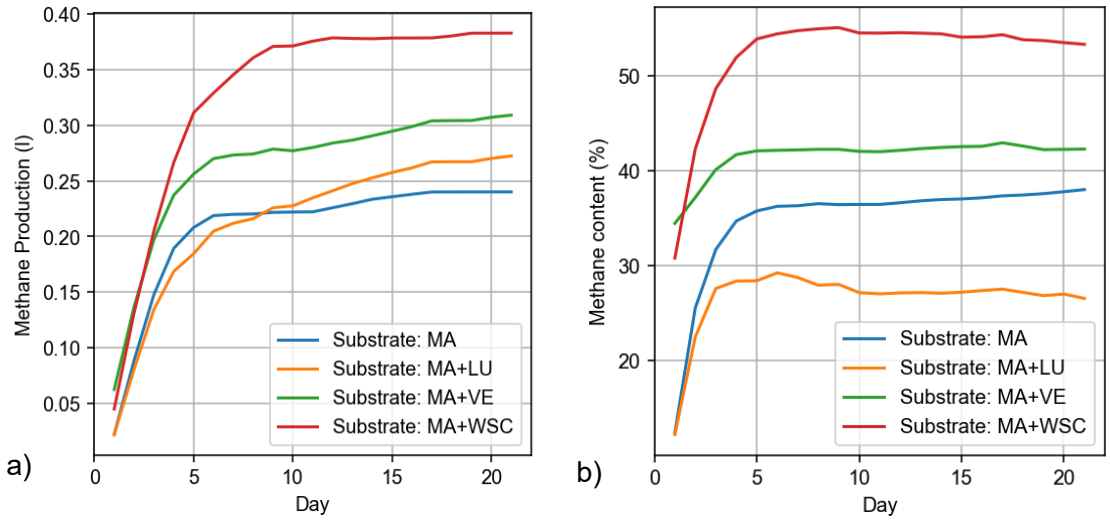

Figure 2. Raw time series from (Kintl et al., 2024) of (a) methane production in liters, and (b) concentration of methane in %.

Features extraction

After applying the one-hot encoding previously discussed in Eqs (1) and (2), we computed the Pearson and Spearman correlation indexes to elucidate which predictor variables can explain the variability of the dependent variables. The results of both tests are depicted in Fig. 3 and Table 3, from which it can be noticed how various parameters of silage composition are related to methane production since the absolute volume of methane (METHANE) and the methane percentage in biogas (METHANE_CONTENT) exhibit high correlations. The strongest positive correlations are between Crude Fiber (CF) and METHANE and METHANE_CONTENT with Pearson correlation coefficients of 0.547 and 0.766, respectively. This would suggest that all of the silages with higher fiber content yield more methane, likely due to these silages having more degradable fibrous material available for microbial degradation. Acid Detergent Fiber (ADF) and Neutral Detergent Fiber (NDF) also show moderately high positive correlations with methane production, once again in agreement with the fact that substrates of fiber lead to greater production of methane.

Furthermore, DAY of sampling has a high correlation with methane formation, i.e., METHANE (0.625 Pearson, 0.634 Spearman). This would be an indication of an increase over time in the formation of methane due to successive degradation of organic matter during fermentation. Its correlation with METHANE_CONTENT is less (0.214 Pearson, 0.194 Spearman), so the amount of methane is rising over time, but its proportion in the gas phase is relatively more stable. Cellulose (CELLULOSIS) and Hemicellulose (HEMI-CELLULOSIS) also correlate positively with methane production at METHANE_CONTENT (0.479 Pearson, 0.471 Pearson for CELLULOSIS and HEMI-CELLULOSIS, respectively). This would imply that these structural carbohydrates are being enriched by methane through their most likely microbial degradation pathways. Conversely, Crude Protein (CP) and Lipids are inversely related to both measures of methane. The crude protein is -0.252 and -0.466

Pearson correlated with METHANE and METHANE_CONTENT, respectively, and lipids have a stronger inhibitory effect on METHANE_CONTENT (-0.324 Pearson, -0.353 Spearman). These patterns suggest that substrates that have greater protein and lipid content may inhibit the production of methane, perhaps due to microbial inhibition or redirection of the metabolic pathways to produce alternative fermentation end products.

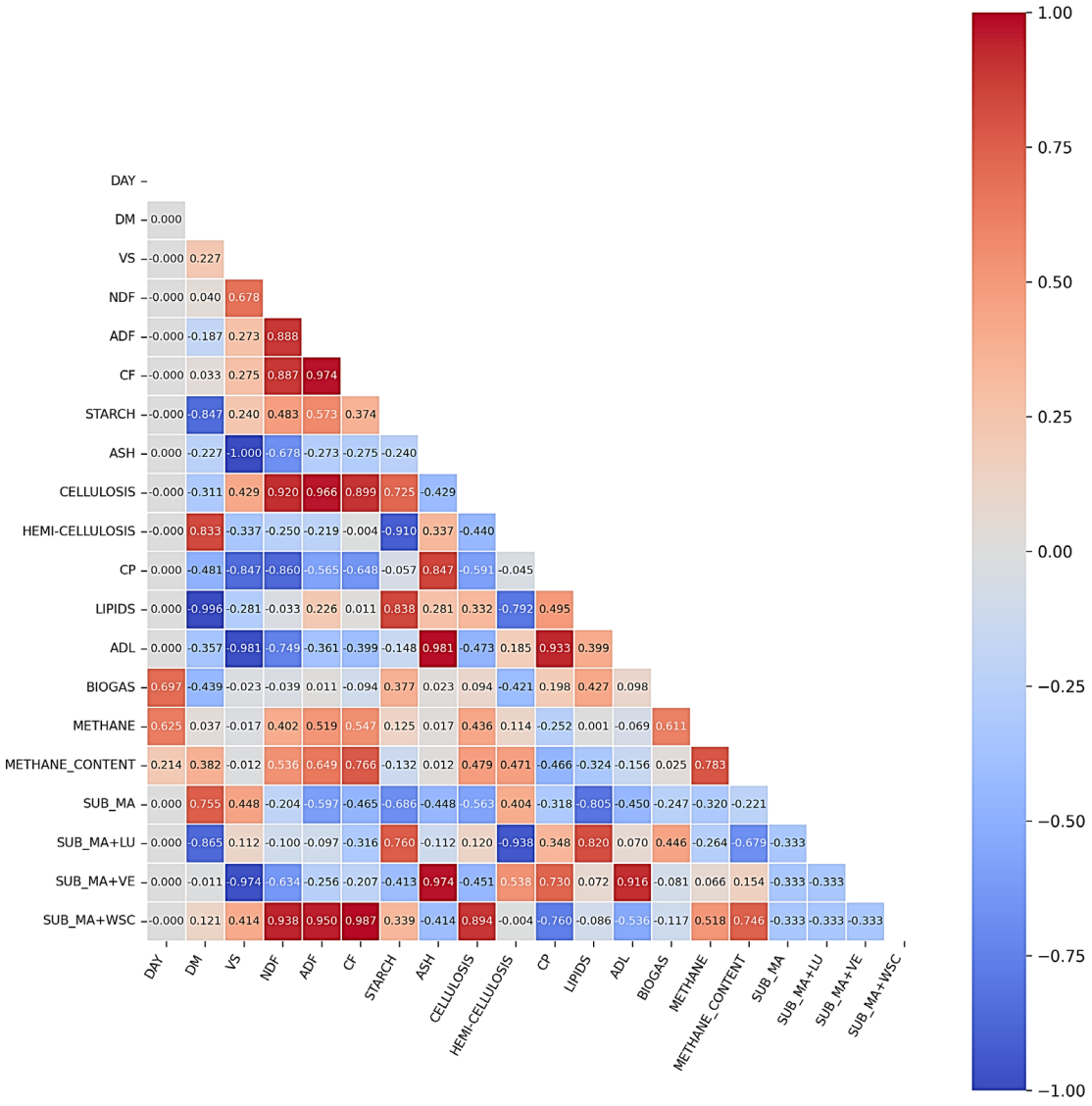

Figure 3. Correlation matrix of the raw data from (Kintl et al., 2024) based on the Pearson index.

Table 3. Pearson and Spearman correlation indexes for METHANE and METHANE_CONTENT versus possible predictor variables of the whole dataset provided by (Kintl et al., 2024). Variables highlighted in bold indicate correlations over 0.5. As a design parameter, we chose these variables since they can explain the variability of the dependent variables over 50%

Predictor variable	Pearson		Spearman	
	METHANE	METHANE CONTENT	METHANE	METHANE CONTENT
DAY	**0.625**	0.214	**0.634**	0.194
DM	0.037	0.382	-0.007	0.353
VS	-0.017	-0.012	-0.191	-0.019
NDF	0.402	**0.536**	0.230	0.169
ADF	**0.519**	**0.649**	0.465	0.363
CF	**0.547**	**0.766**	**0.649**	**0.735**
STARCH	0.125	-0.132	0.144	-0.181
ASH	0.017	0.012	0.191	0.019
CELLULOSIS	0.436	0.479	0.465	0.363
HEMI-CELLULOSIS	0.114	0.471	0.090	0.375
CP	-0.252	-0.466	-0.180	-0.346
LIPIDS	0.001	-0.324	0.007	-0.353
ADL	-0.069	-0.156	-0.180	-0.346
SUB_MA	-0.320	-0.221	-0.424	-0.247
SUB_MA+LU	-0.264	**-0.679**	-0.293	**-0.706**
SUB_MA+VE	0.066	0.154	0.182	0.255
SUB_MA+WSC	**0.518**	**0.746**	**0.535**	**0.698**

The correlations also indicate that silage compositions incorporating white sweet clover (MA+WSC) exhibit the strongest positive correlation with METHANE_CONTENT (0.746 Pearson, 0.698 Spearman), suggesting that this mixture enhances methane enrichment in the produced biogas. This could be attributed to the biochemical composition of white sweet clover, which may provide optimal fermentable substrates for methanogenesis. In contrast, silages containing white lupin (MA+LU) show a strong negative correlation with METHANE_CONTENT (-0.679 Pearson, -0.706 Spearman), indicating that its inclusion may reduce methane concentration in the biogas. The high protein content and potential presence of bioactive compounds in white lupin might inhibit methanogenic activity, leading to lower methane yields. The impact of fodder vetch (MA+VE) appears to be more neutral, with weak positive correlations (0.066 Pearson, 0.255 Spearman), suggesting that while it does not significantly enhance methane concentration, it also does not strongly inhibit it. These findings imply that white sweet clover is the most favorable silage component for maximizing methane content, while white lupin may hinder methane production, potentially making it less desirable for biogas optimization. In summary, the exogenous variables for the ARIMAX models for METHANE are ADF, CF, and SUB_MA+WSC. Regarding the exogenous variables for METHANE_CONTENT, we selected NDF, ADF, CF, SUB_MA+LU, and SUB_MA+WSC.

Regarding the use of ARIMAX models, we have to assess the following conditions: *i*) determine if the METHANE and METHANE_CONTENT series are stationary using the Augmented Dickey-Fuller (ADF) test (Eqs (4) and (5)),

ii) differentiate the series until stationarity (Eq. (6)), and *iii)* use the ACF and PACF to elucidate which lags and past errors contribute to the predictions. For the first and second steps, we conducted the ADF tests, whose results are depicted in Table 4. From this, it can be noticed that the differentiation of the time series achieved stationarity since the ADF statistics are greater than the critical values and the *p*-values are less than 0.05, equivalent to a statistical significance of 95%. In that way, we chose $d = 1$ in the ARIMAX model; that is, we will use the first difference instead of the raw time series.

Table 4. Augmented Dickey-Fuller Test for METHANE and METHANE_CONTENT time series without differentiation and with differentiation. If the ADF statistic is greater than the critical value or the *p*-value is less than 0.05 (95% confidence), the resulting time series is stationary

Differentiation order	METHANE			METHANE_CONTENT			Conclusion
	ADF statistic	Critical value	*p*-value	ADF statistic	Critical value	*p*-value	
No differentiation	-2.416	-4.137	0.137	0.757	-3.889	0.99	No stationary
Differentiation 1	-4.58	-3.517	0.0001	-6.862	-3.516	1.59E-09	Stationary

For the third step, we computed the ACF and PACF to elucidate which lags and past errors should be considered for the corresponding models using the differentiated series. The plots of these functions are depicted in Fig. 4. Each vertical line indicates the lag in the case of the ACF and the past error in the case of the PACF. If a line is over the blue-shaded region, it means that the corresponding lag/error should be considered for the regression.

From Fig. 4, (a), it can be noticed that the plot of the ACF of the differenced METHANE series is positive at lag 0 and negative at lag 1, and the remaining values fluctuate within the confidence band (blue-shaded area). It shows that the moving average term has to include error terms at lag 0 and 1 and because the subsequent lags do not show any correlation. Besides, from Fig. 4, (b), it can be noticed that the PACF also shows that the lags 0, 1, and 2 are significant, indicating that an autoregressive term of order 3 would be sufficient to capture dependencies in the data. These patterns indicate that an ARIMAX model for METHANE would consist of three autoregressive terms, two moving average terms, and the exogenous variables ADF, CF, and SUB_MA+WSC, that is, an ARIMAX(3,1,2) expressed as

$$\mathrm{CH4}_t = \beta_0 + \beta_1 \mathrm{ADF} + \beta_2 \mathrm{CF} + \beta_3 \mathrm{WSC} + \phi_1 \mathrm{CH4}_{t-1} + \phi_2 \mathrm{CH4}_{t-2} + \phi_3 \mathrm{CH4}_{t-3} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}, \quad (8)$$

where $\mathrm{CH4}_t$ represents the methane volume at the time *t,* ADF represents the Acid Detergent Fiber content, CF represents the Crude Fiber content, WSC represents the silage with maize and white sweet clover, $\mathrm{CH4}_t$ represents the methane volume in the previous day, $\mathrm{CH4}_{t-2}$ represents the methane volume two days ago, β_0 represents the independent term, β_1 is the weight for the ADF, β_2 is the weight for the CF, β_3 is the weight for the silage prepared with maize and white sweet clover WSC, ϕ_1 is the weight for the lag 1 of CH4, ϕ_2 is the weight for the lag 2 of CH4, ϕ_3 is the weight for the lag 3 of CH4, θ_1 is the weight for the lagged error ϵ_{t-1}, and $\theta_2$is the weight for the lagged error ϵ_{t-2}.

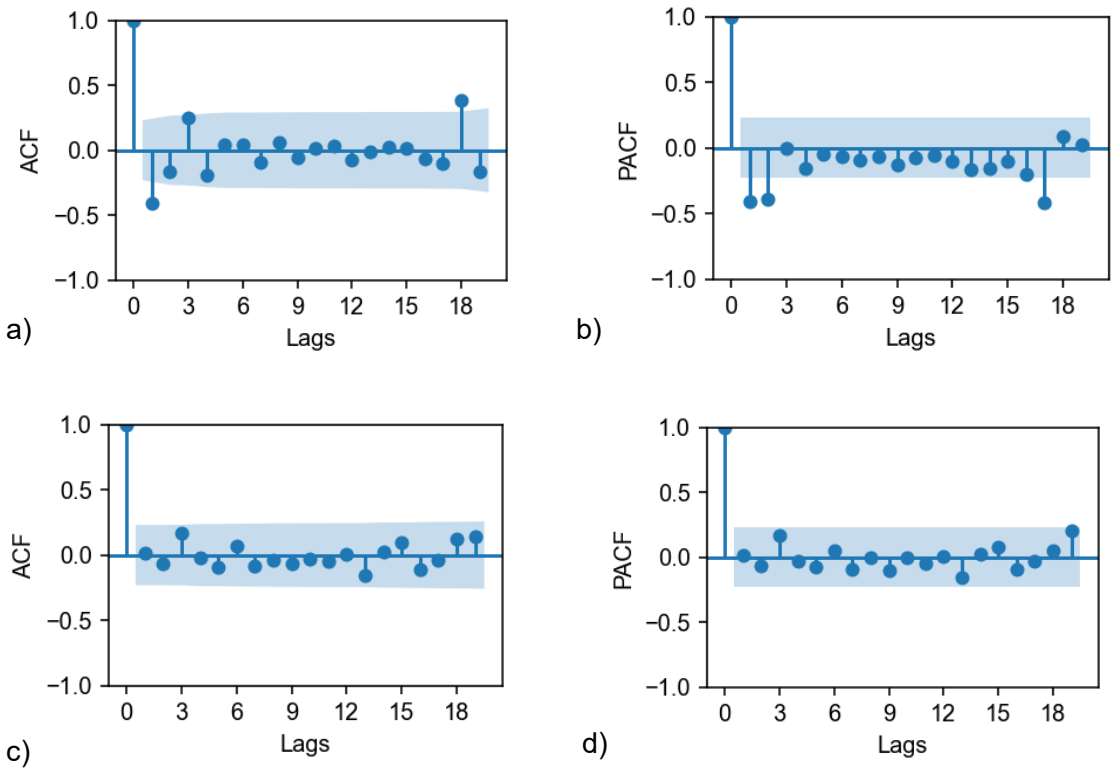

Figure 4. Plots of (a) ACF of METHANE, (b) PACF of METHANE, (c) ACF of METHANE_CONTENT, and (d) PACF of METHANE_CONTENT. The plots are for the differentiated series.

Regarding the METHANE_CONTENT, we analyze Fig. 4, (c), and (d) where the following can be noticed. First, Fig. 4 (c), shows the ACF plot and indicates that only the lag 0 is significant (the first difference), so it suggests that only the first lagged error should be included in the model because the other lags are inside the blue-shaded region. Besides, Fig. 4, (d), shows the PACF, suggesting the same configuration, i.e., only the first lagged difference should be considered. This implies that a single moving average and a single autoregressive term would be appropriate. Since both series required one difference to become stationary, an ARIMAX(1,1,1) specification is a feasible model of the underlying dynamics described by:

$$\text{C\_CH4}_t = \beta_0 + \beta_1 \text{ADF} + \beta_2 \text{CF} + \beta_3 \text{WSC} + \beta_4 \text{NDF} + \beta_5 \text{LU} + \phi_1 \text{C\_CH4}_{t-1} + \theta_1 \epsilon_{t-1}, \quad (9)$$

where C_CH4_t represents the methane content at the time t, CH4_{t-1} represents the methane content in the previous day, NDF represents the Neutral Detergent Fiber, LU represents the silage with maize and white lupin, β_0 represents the independent term, β_1 is the weight for the ADF, β_2 is the weight for the CF, β_3 is the weight for the silage prepared with maize and white sweet clover WSC, β_4 is the weight for NDF, β_5 is the weight for the silage prepared with maize and white lupin, ϕ_1 is the weight for the lag 1 of C_CH4, and θ_1 is the weight for the lagged error ϵ_{t-1}.

Modeling

This subsection shows the fitted models and the corresponding results in predicting METHANE and METHANE_CONTENT. We also provide a goodness-of-fit analysis for each fitted model.

ARIMAX model for METHANE

We fitted the model in Eq. (8), obtaining the coefficients in Table 5. To carry out the fitting process, we used the first 16 days of METHANE values and the values from 17 to 21 to perform the goodness-of-fit analysis. We aim to evaluate how the models perform with unknown data that has not been observed previously in the training phase.

The coefficients in Table 5 help analyze the contribution of different factors to methane production from physical and statistical standpoints. The variables of substraate composition, ADF, CF, and WSC have coefficients expressed as liters per percentage unit or liters and indicate how the methane yield is affected by variation in the feedstock composition. The moving average components (θ_i) and autoregressive components (ϕ_i) are adimensional because they measure the impact of past values of methane and past errors of forecasting on the observed output. Furthermore, we can analyze the magnitudes and signs of the coefficients to draw conclusions about the effects of the different variables. For instance, the Acid Detergent Fiber coefficient is 0.0697 liters per unit percentage,or a 1% increase in ADF causes an increase of 0.0697 liters of methane. This relatively low value suggests that ADF exerts a very small effect on methane production, most likely because it is composed of cellulose and lignin, which are broken down slowly. For the case of Crude Fiber (CF), the model obtains a more pronounced effect with a coefficient of 0.1327 liters per percentage unit, which implies that substrates with more crude fiber produce more methane. Besides, this suggests that CF is a more effective predictor of methane yield than ADF. Finally, the coefficient for WSC (0.6045 liters) showed the largest contribution, which indicates that the availability of this substrate significantly triggers methane production.

Table 5. Fitted weights for the ARIMAX(3,1,2) model for METHANE

Coefficient	Variable	Value	Units
β_1	ADF	0.0697	liter/%
β_2	CF	0.1327	liter/%
β_3	WSC	0.6045	liter
ϕ_1	$CH4_{t-1}$	0.2016	Adim
ϕ_2	$CH4_{t-2}$	-0.0402	Adim
ϕ_3	$CH4_{t-3}$	-0.427	Adim
θ_1	ϵ_{t-1}	0.6417	Adim
θ_2	ϵ_{t-2}	-0.413	Adim

On the other hand, the autoregressive coefficients (ϕ_1, ϕ_2, ϕ_3) capture the impact of past methane values on current methane output. The positive value of ϕ_1 at 0.2016 indicates that an increase in methane production at the time *t*-1 causes an increase in methane at the current time *t*, indicating some persistence in methane production. However, the negative values for ϕ_2 = -0.0402 and ϕ_3 = -0.427 imply a suppression effect with lag, where methane levels two and three times in the past contributed to lowering current methane output. The more negative value of ϕ_3 indicates that this oscillatory effect is stronger after three-time steps, perhaps due to substrate depletion, microbial adaptation, or inhibitory compounds affecting methane production in the long term.

The parameters of the moving average (θ_1, θ_2) capture the effect of past errors in predicting the current methane production. The positive value of $\theta_1 = 0.6417$ means that higher-than-average methane production in the previous time step increases the current value, and this helps support short-term fluctuations. The negative value of $\theta_2 = -0.413$ means that two periods' past errors work in the opposite way and counter or reduce very large fluctuations.

Regarding the accuracy of the ARIMAX model for METHANE, we obtained an R^2 of 0.92 and an RMSE of 0.001 liters, which demonstrates the ability of the fitted model to forecast the methane volume production when different exogenous and time series-based variables are considered. To show graphically the quality of the forecasts provided by the ARIMAX model, we present Fig. 5 from the following can be noticed.

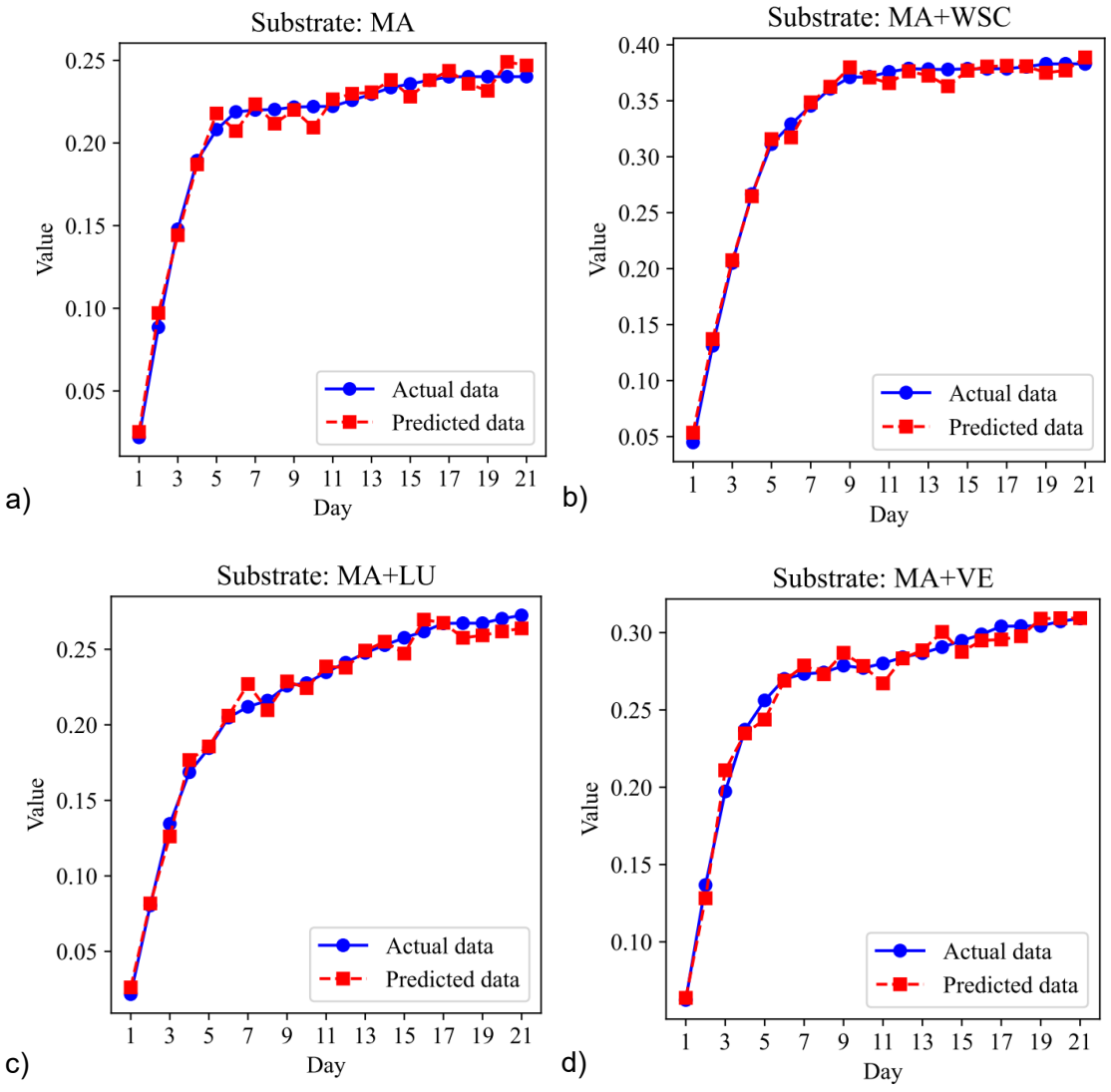

Figure 5. Performance of the ARIMAX model for METHANE for (a) maize, (b) maize plus white sweet clover silage, (c) maize plus white lupin silage, and (d) maize plus fodder vetch silage.

The forecast accuracy (red dashed line with square markers) adequately captures the methane production as time increases. The forecasted values closely follow the real data trend, accurately catching the initial increase as well as the later stabilization phases. The predicted and actual values show low deviations, particularly in the initial and mid-periods of methane production. However, there are some minor deviations at later stages since the calculated values occasionally fall below or above the measured values. Despite these minor differences, the model can create the overall trend and show its high predictive power.

ARIMAX model for METHANE_CONTENT

We fitted the model in Equation (9), obtaining the coefficients in Table 6. . To carry out the fitting process, we used the same process of separating the first 16 days of METHANE_CONTENT values and used the values from 17 to 21 to evaluate how the model performs with unknown data.

Table 6. Fitted weights for the ARIMAX(1,1,1) model for METHANE

Coefficient	Variable	Value	Units
β_1	ADF	0.1823	Adim
β_2	CF	0.2532	Adim
β_3	WSC	0.3815	%
β_4	NDF	0.0824	Adim
β_5	LU	-0.5245	%
ϕ_1	C_CH4_{t-1}	0.3241	Adim
θ_1	ϵ_{t-1}	0.2841	Adim

Physically, the coefficients in Table 6 define the effect of different factors on methane concentration dynamics. For instance, the coefficient for ADF ($\beta_1 = 0.1823$) shows that acid detergent fiber has a positive effect on methane production at a moderate level. ADF is the least digestible fraction of fiber, primarily cellulose and lignin. Its positive effect shows that although it is not fully degradable, there is still a portion that is fermented by microbes, producing methane. The coefficient for CF ($\beta_2 = 0.2532$) indicates that crude fiber has a more important positive effect compared to ADF. Crude fiber consists of cellulose and a fraction of the hemicelluloses, which are fermented in part by methanogenic microorganisms. This indicates that the digestibility of the fiber component plays a key role in the methane concentration. With respect to the coefficient for WSC ($\beta_3 = 0.3815$), it can be noticed that the use of a combination of maize and white sweet clover has a significant positive effect on regulating methane.Since WSC includes fermentable sugars, which are readily fermented to volatile fatty acids and, eventually, methane in anaerobic conditions, this is directed towards the key driving role of fermentable substrates in methane emission. Conversely, the coefficient for NDF ($\beta_4 = 0.0824$), which symbolizes neutral detergent fiber, exhibits a significantly smaller positive effect. This small effect means that while some part of its composition is responsible for the generation of methane, the NDF does not allow microbial fermentation to produce high concentrations of methane. Furthermore, the coefficient for LU ($\beta_5 = -0.5245$) is negative, and the suggestion is that the addition of white lupin inhibits the methane concentration increase. This inhibition suggests that white lupin constrains microbial activity, reduces the level of fermentable substrates, or disrupts the anaerobic digestion.

The autoregressive coefficient $\phi_1 = 0.3241$ indicates that a previous methane content significantly affects the current values. This means that the methane concentration has an autoregressive pattern, which holds that environments that are supportive of methane creation in earlier stages are likely to maintain production for

longer periods. The coefficient for the moving average $\theta_1 = 0.2841$ explains how past random errors impact the current methane concentration. It indicates that a variation in methane generation is explained in part by omitted variables in the last period.

Regarding the accuracy of the ARIMAX model for METHANE_CONTENT, we obtained an R^2 of 0.908 and an RMSE of 0.85, which demonstrates the fitted model's ability to forecast the methane concentration with the lagged variables, errors, and exogenous variables chosen. We depict the actual and predicted trends in Fig. 6.

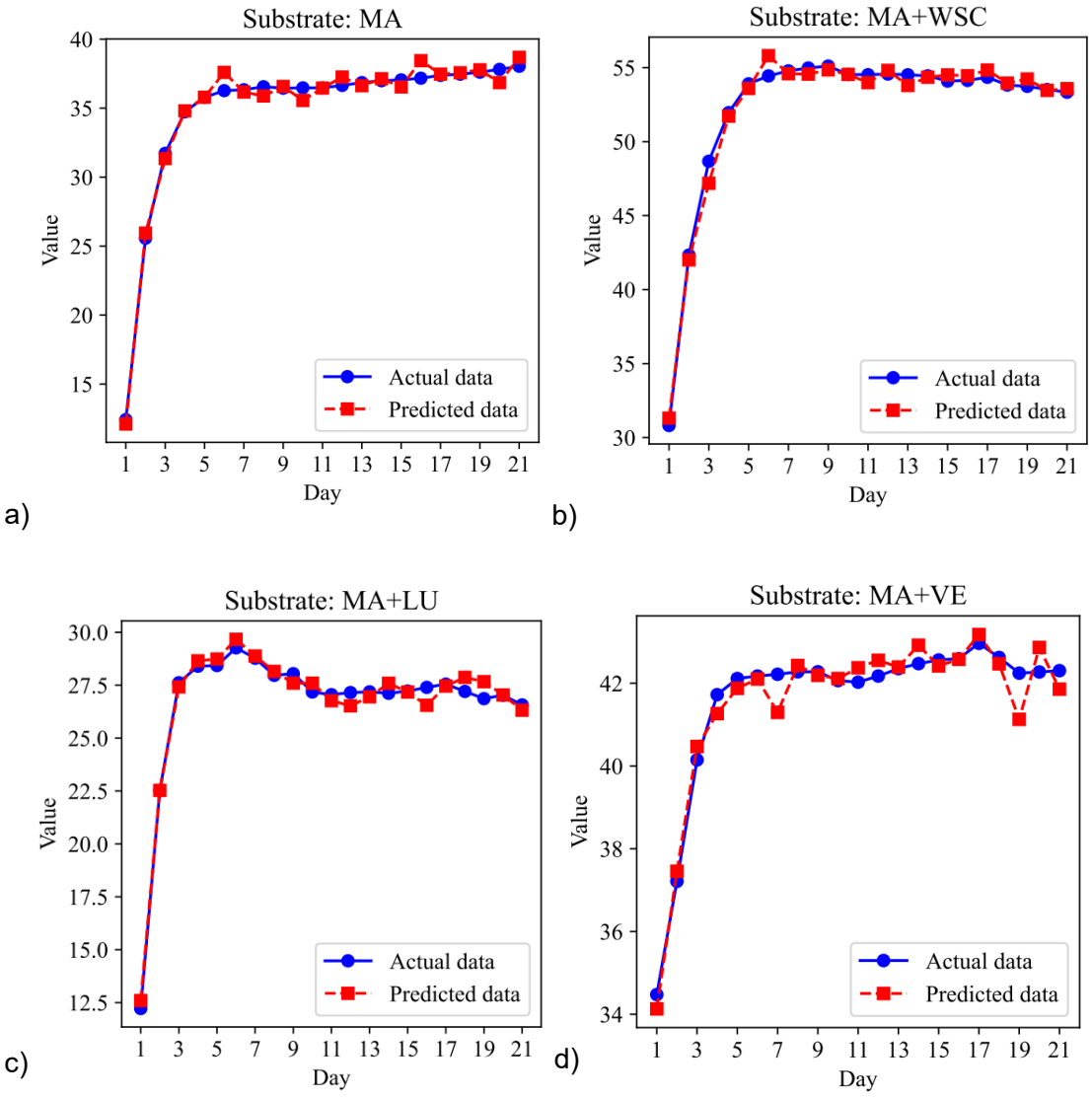

Figure 6. Performance of the ARIMAX model for METHANE_CONTENT for (a) maize, (b) maize plus white sweet clover silage, (c) maize plus white lupin silage, and (d) maize plus fodder vetch silage. The y-axis is in units of % of methane.

From Fig. 6, the following can be observed. First, it can be noticed that the ARIMAX model has a strong predictive power for the considered substrates, as observed from the very close convergence of true and forecasted values. The model captures the

sharp early increase in methane production accurately and also the subsequent stabilization. Based on y-axis values, white sweet clover + maize (MA+WSC) possesses the maximum methane content with values of more than 55%. This verifies the hypothesis that there are readily fermentable substrates present in white sweet clover, through which there can be rapid microbial degradation leading to the generation of methane. For MA only, methane content levels are at about 38-40%, with a slightly lower terminal concentration than MA+WSC. Despite this difference, the model is highly predictive in accuracy, tracking the actual path very closely with little variation. For the case of MA+LU silage, a fluctuating trend can be noticed, a methane peak of approximately 30%, with a subsequent drop, and then a flat line. In all the cases, the prediction accuracy of the ARIMAX model suggests an adequate ability to forecast the methane concentration under different conditions of exogenous variables.

With these results, we show that our ARIMAX models exhibit an adequate predictive ability (an R^2 of 0.92 and an RMSE of 0.001 L for methane volume, and an R^2 of 0.908 with 0.85% RMSE for methane concentration), indicating that silage composition and its time variation can predict gas outputs with confidence. In practice, this means that operators can use these models for real-time process control, adjusting feedstock mix (for example, favoring maize + white sweet clover, which our forecasts show yields maximum methane concentrations above 55%) in an effort to maximize biogas quality and quantity. Most significantly, the coefficients associated with the exogenous variables (e.g., CF's 0.1327 L/% and WSC's 0.6045 L unit change) directly translate to operational recommendations on how most strongly individual silage components affect the creation of methane, allowing decision-making regarding crop mix and settings from evidence to farmers and managers.

CONCLUSIONS

Generation of biogas from mixed maize-legume silages is a renewable-energy source that is sustainable, but its wider adoption may have been prevented by uncertainty about the quantitative value of different mixtures of feeds. To bridge this knowledge gap, we developed a time-series-based forecasting model founded on parametric ARIMAX(p,d,q) techniques that aim to explain a series' internal dynamics and exogenous silage chemistry variables, which can be selected through Pearson and Spearman correlation analyses. By translating categorical types of substrates into one-hot encoded representations and choosing the most effective predictors, our methodology empowers statistical integrity with functional usability.

Our feature-selection analysis showed that acid detergent fiber (ADF) and crude fiber (CF) are the most important positive predictors of methane yields. The CF contributes to biogas volume, confirming the principal role of digestible fiber in facilitating sustained methane production. The ADF contributes a smaller effect, as would be expected because of its lower rates of digestion in the digester. Conversely, a rise in crude protein and lipid levels inhibits methane production, most likely due to ammonia accumulation and long-chain fatty-acid toxicity, directing microbial metabolism towards non-methanogenic products.

The ARIMAX(3,1,2) model for methane volume yielded an R^2 of 0.92 and RMSE of 0.001 L, while the ARIMAX(1,1,1) model for methane concentration achieved an R^2 of 0.908 with RMSE of 0.85%. Unlike black-box regressors such as LSTM networks, which provide minimal insight into the effect of individual feedstock components on performance, our transparent ARIMAX approach offers insightful coefficient estimations. This transparency allows operators to make real-time silage blend adjustments to maximize biogas yield and quality using evidence-based process control.

USE OF GENERATIVE AI DECLARATION: The authors confirm that they used ChatGPT to fix typo errors, improve grammar, translate some excerpts from Spanish, and enhance the readability of the manuscript. However, they declare that the manuscript is original, and the improved texts obtained from the language models have been carefully reviewed.

REFERENCES

Avinash, L.S. & Mishra, A. 2024. Comparative evaluation of artificial intelligence based models and kinetic studies in the prediction of biogas from anaerobic digestion of MSW. *Fuel* **367**, 131545.

Farzin, F., Moghaddam, S.S. & Ehteshami, M. 2024. Auto-tuning data-driven model for biogas yield prediction from anaerobic digestion of sewage sludge at the South-Tehran wastewater treatment plant: Feature selection and hyperparameter population-based optimization. *Renewable Energy* **227**, 120554.

Ferreira, J., Santos, L., Ferreira, M., Ferreira, A. & Domingos, I. 2024. Environmental assessment of pig manure treatment systems through life cycle assessment: A mini-review. *Sustainability* **16**(9), 3521.

Geng, Z., Shi, X., Ma, B., Chu, C. & Han, Y. 2024. Biogas production prediction model of food waste anaerobic digestion for energy optimization using mixup data augmentation-based global attention mechanism. *Environmental Science and Pollution Research* **31**(6), 9121–9134.

Gonzalez-Palacio, M., Moncada, S.V., Luna-delRisco, M., Gonzalez-Palacio, L., Montealegre, J.J.Q., Orozco, C.A.A., Diaz-Forero, I., Velasquez, J.-P. & Marin, S.-A. 2018. Internet of things baseline method to improve health sterilization in hospitals: An approach from electronic instrumentation and processing of steam quality. In: *Proc. 13th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6.

González-Palacio, M., Tobón-Vallejo, D., Sepúlveda-Cano, L.M., Luna-delRisco, M., Roehrig, C. & Le, L.B. 2024. Machine-learning-assisted transmission power control for LoRaWAN considering environments with high signal-to-noise variation. *IEEE Access* **12**, 54449–54470. doi: 10.1109/ACCESS.2024.3387457

Himmel, M.E., Ding, S.-Y., Johnson, D.K., Adney, W.S., Nimlos, M.R., Brady, J.W. & Foust, T.D. 2007. Biomass recalcitrance: Engineering plants and enzymes for biofuels production. *Science* **315**(5813), 804–807.

Hoang, A.T., Goldfarb, J.L., Foley, A.M., Lichtfouse, E., Kumar, M., Xiao, L., Ahmed, S.F., Said, Z., Luque, R. & Bui, V.G. 2022. Production of biochar from crop residues and its application for anaerobic digestion. *Bioresource Technology* **363**, 127970.

Kaparaju, P. & Rintala, J. 2005. Anaerobic co-digestion of potato tuber and its industrial by-products with pig manure. *Resources, Conservation and Recycling* **43**(2), 175–188.

Kim, H.G., Yu, S.Il, Shin, S.G. & Cho, K.H. 2025. Graph-based deep learning for predictions on changes in microbiomes and biogas production in anaerobic digestion systems. *Water Research* 123144.

Kintl, A., Huňady, I., Sobotková, J., Vítěz, T., Brtnický, M., Vejražka, K. & Elbl, J. 2024. Data on the effect of co-fermentation of maize and leguminous crops on biogas production, methane production and methane content in biogas. *Data in Brief* **56**, 110842.

Liu, X., Qi, L., Chatzisymeon, E., Yang, P., Sun, W. & Pang, L. 2021. Inorganic additives to increase methane generation during anaerobic digestion of livestock manure: A review. *Environmental Chemistry Letters* **19**, 4165–4190.

Mamidala, S., Mohan, G. & Veeramani, C. 2023. Hybrid renewable energy resources accuracy, techniques adopted, and the future scope abetted by the patent landscape–A conspicuous review. *Energy Harvesting and Systems* **10**(2), 213–228.

Mao, C., Feng, Y., Wang, X. & Ren, G. 2015. Review on research achievements of biogas from anaerobic digestion. *Renewable and Sustainable Energy Reviews* **45**, 540–555.

Okolie, J.A., Jimoh, T., Akande, O., Okoye, P.U., Ogbaga, C.C., Adeleke, A.A., Ikubanni, P.P., Güleç, F. & Amenaghawon, A.N. 2023. Pathways for the valorization of animal and human waste to biofuels, sustainable materials, and value-added chemicals. *Environments* **10**(3), 46.

Ologunde, O.H., Akanni, S.D., Olayemi, A.B. & Busari, M.A. 2025. Assessment of simple engineering approaches and poultry manure for soil erosion control under maize cultivation in the tropics. *Air, Soil and Water Research* **18**, 11786221241311724

Palacio, M.G., Palacio, L.G., Montealegre, J.J.Q., Pabón, H.J.O., Del Risco, M.A.L., Roldán, D., Salgarriaga, S., Vásquez, P., Hernández, S. & Martínez, C. 2017. A novel ubiquitous system to monitor medicinal cold chains in transportation. In: *Proc. 12th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6.

Pfordt, A. & Paulus, S. 2025. A review on detection and differentiation of maize diseases and pests by imaging sensors. *Journal of Plant Diseases and Protection* **132**(1), 1–21.

Prasad, N.R.C.S., Ganeshan, P. & Rajendran, K. 2023. Forecasting of biogas production using advanced time series algorithms 2.

Thakur, N., Sharma, M., Alghamdi, H., Zheng, Y., Xue, W., Jeon, B.-H., Salama, E.-S. & Li, X. 2023. A recent trend in anaerobic digestion (AD): Enhancement of microbiome and digestibility of feedstocks via abiotic stress factors for biomethanation. *Chemical Engineering Journal* 145047.

Tiismus, H., Maask, V., Astapov, V., Korõtko, T. & Rosin, A. 2025. State-of-the-art review of emerging trends in renewable energy generation technologies. *IEEE Access* **13**, 10820–10843. https://doi.org/10.1109/ACCESS.2025.3528640

Zou, J., Lü, F., Chen, L., Zhang, H. & He, P. 2024. Machine learning for enhancing prediction of biogas production and building a VFA/ALK soft sensor in full-scale dry anaerobic digestion of kitchen food waste. *Journal of Environmental Management* **371**, 123190.