

Genetic Algorithms and Multiple Regression : A Statistical Comparison for Parameter Identification of Biogas Production Model via Null Hypothesis Test

Abdelhani CHAABNA¹, Samia SEMCHEDDINE¹

¹Power Electronics and Industrial control Laboratory (LEPCI),
Faculty of Technology, University of Sétif-1, Sétif 19000, Algeria

E-mail: abdelhani.chaabna@gmail.com

Abstract - — The generation of biogas through anaerobic digestion serves a dual purpose. It not only enables the creation of sustainable energy but also aids in the management and disposal of organic waste. Parameter identification is a crucial step in modeling and control systems. Most estimates are reported in the literature without any analysis of uncertainty. This paper presents a comparative study of two parameter estimation techniques: Multiple Linear Regression (MLR) and Genetic Algorithms (GA). These techniques were employed first in parameter identification of a reduced biogas production model. The p-value was then calculated for the obtained estimates. The analysis was performed by testing the null hypothesis. Therefore, it is the accuracy assessment of the obtained estimates. Simulations were executed on Matlab.

Keywords – Parameter identification, P-value, Genetic algorithm, Multiple linear regression, biogas production model, Null hypothesis.

I. INTRODUCTION

Anaerobic digestion (AD) is a biotechnological process that plays a crucial role in numerous life processes. It's seen as a potential solution to various energy and environmental challenges faced by agriculture and agro-industry. This process involves the decomposition of biodegradable materials (organic matter) by microorganisms in an oxygen-free environment, resulting in the production of biogas. This biogas primarily consists of methane and carbon dioxide, along with traces of other gases. Continuously stirred tank bioreactors (CSTR) are commonly used for this purpose.

Mathematical models serve as a means to depict the primary characteristics of a biological system. They enhance our comprehension of the system, assist in formulating and validating hypotheses, and predict the system's response under varying conditions. As a result, they help reduce the need for experimental data, thereby saving costs, risks, and time. For mathematical models to be effectively utilized in bioprocesses,

several steps must be followed. The ultimate aim of this approach is to develop useful tools that can enhance the system.

Upon establishing the modeling objectives and gathering data, two key considerations come into play. The first consideration is whether the chosen model structure has the capability to accommodate the measured data. This requires the model to possess the necessary degrees of freedom, but not excessively so, as it could lead to over-parametrization [1]. The second consideration is whether it's feasible to identify a unique optimal set of parameters based on the experimental data available, once a model structure has been selected.

AD models, in essence, are mechanistic models that encapsulate a vast amount of scientific research aimed at comprehending the physical, chemical, and biological mechanisms of the processes involved. As a result, most model parameters have a physical significance and default values are typically provided [2]. The issue of identifiability is a delicate one where the modeler should only calibrate those parameters necessary to explain the observed mechanisms

without overfitting the data. In other words, an overcalibrated model may reproduce the experimental data quite well but would lose its predictive or exploratory capabilities.

Moreover, in industrial applications, parameter identification is crucial as well. It forms the basis for model development and optimization, which are key to predicting and controlling various processes [3,4].

Bioprocess models are inherently complex and nonlinear, consisting of a combination of differential and algebraic equations. A critical aspect of applying these models is identifying an appropriate set of parameters that accurately represent the system under study. However, the assumptions made during model formulation, along with process variability and measurement errors, introduce uncertainty into the estimated parameters. This uncertainty is an important factor that needs to be quantified and accounted for in the modeling process.

Recently, there has been a concerted effort to account for model uncertainty during the optimization and control of bioprocesses, as evidenced by the work of [5]. These advancements underscore the importance of quantifying the uncertainty of a kinetic model for its practical application in process design. It is also highly recommended to employ hypothesis testing to scrutinize and justify the inclusion of each parameter in a kinetic model, ensuring that its estimated value is statistically significant. Furthermore, hypothesis testing should be utilized when comparing estimated values for a parameter derived from different experiments.

Iweka et al. [6] used Python to statistically analyze the production of biogas from co-digestion of corn chaff and cow dung. They obtained a P-value less than 0.0001, which indicates that the relationship between the variables is significant. This means that the null hypothesis, which states that there is relationship between the variables, is rejected.

Taus et al. used statistical analysis to evaluate the suitability of kitchen waste and agricultural crops as feedstocks for biogas production [7]. A total of 854 data points were analyzed using a one-way Analysis of variance (ANOVA) test.

The results showed that the individual substrate had a significant effect on methane production, with a P-value less than 0.0001.

This paper presents a novel comparison of the performance of two parameter identification techniques (genetic algorithm and multiple linear regression) for methane production using the p-value analysis. The prediction accuracy of the two techniques is evaluated using observed data from the ADM1.

II. MATERIALS AND METHODS

The parameters of AM2HN model was identified by MLR in [8], and a comparison of the MLR and by genetic algorithm was released in [9], the AM2HN ODEs system is given by:

$$\frac{dX_T}{dt} = D(X_{T,in} - X_T) - k_{hyd}X_T \quad (1)$$

$$\frac{dX_1}{dt} = (\mu_1(S_1) - \alpha D)X_1 \quad (2)$$

$$\frac{dX_2}{dt} = (\mu_2(S_2) - \alpha D)X_2 \quad (3)$$

$$\frac{dS_1}{dt} = D(S_{1,in} - S_1) - k_1\mu_1(S_1)X_1 + k_{hyd}X_T \quad (4)$$

$$\frac{dS_2}{dt} = D(S_{2,in} - S_2) + k_2\mu_1(S_1)X_1 - k_3\mu_2(S_2)X_2 \quad (5)$$

$$\frac{dC}{dt} = D(C_{in} - C) - q_c + k_4\mu_1(S_1)X_1 + k_5\mu_2(S_2)X_2 \quad (6)$$

Where alpha is the biomass retention coefficient. C represents the total inorganic carbon concentration, measured in kmoleCm⁻³. D is the dilution rate coefficient, measured in d⁻¹. k_1 to k_6 , and k_{hyd} are yield coefficients for various processes. k_i , k_{S1} , and k_{S2} are inhibition and half saturation constants, measured in KgCODm⁻³. q_c and q_{ch4} are the flow rates of carbon dioxide and methane respectively. S_1 and S_2 represent the concentrations of organic substrate and volatile fatty acids respectively, measured in KgCODm⁻³. X_1 and X_2 are the concentrations of acidogenic and methanogenic bacteria respectively, measured in KgCODm⁻³. $\mu_1(S_1)$, μ_{1max} , $\mu_2(S_2)$, and μ_{2max} are specific growth rates and maximum growth rates of acidogenic and methanogenic bacteria respectively, measured in d⁻¹. $X_{T,in}$, $S_{1,in}$, $S_{2,in}$, and C_{in} are input values for total particulate substrate, organic substrate concentration, volatile fatty acids concentration, and total

inorganic carbon concentration respectively. The growth rate of acidogenic bacteria $\mu_1(S_1)$, follows a Monod model:

$$\mu_1(S_1) = \frac{\mu_{1\max} \times S_1}{S_1 + k_{S1}} \quad (7)$$

The growth rate of methanogenic bacteria $\mu_2(S_2)$, follows a Haldane model:

$$\mu_2(S_2) = \frac{\mu_{2\max} \times S_2}{S_2 + k_{S2} + S_2^2/k_i} \quad (8)$$

III. DATA EXTRACTION

The steady-state data used in this study were adapted from Table 3 of [9]. These data were produced by simulating the single-stage mesophilic anaerobic digestion (AD) of waste activated sludge in a continuous stirred-tank reactor (CSTR) without biomass retention at different dilution rates.

IV. RESULTS AND DISCUSSION

In the context of our study, we employed two distinct methodologies as shown in [9]: Multiple Linear Regression (MLR) and a Genetic Algorithm (GA). The table 1 and table 2 represent the residuals related to the MLR model and genetic algorithm model respectively.

A) Genetic algorithm

Genetic Algorithms (GA) have been in existence for several decades, originating from the influential work of Goldberg in 1989 [11]. They have been predominantly utilized in the field of optimization, as documented by Michalewicz Z [12]. These algorithms are probabilistic in nature and are based on the principle of Artificial Darwinism. This principle asserts that only the fittest individuals survive, leading to the evolution of the species [10].

Table 1. Matlab configuration of GA

Ga parameter	Value/Function
Selection function	Stochastic uniform
Population size	200
Generation size	700
Mutation function	Adaptive feasible
Crossover	Scattered
Tolerance function	1e-6

Table 1 outlines the Matlab configuration of the parameters for a GA. The selection function is set to 'Stochastic uniform', indicating that individuals are selected from the population randomly but with a probability proportional to their fitness. The population size is 200, meaning there are 200 potential

solutions in each generation. The algorithm will run for 700 generations.

The mutation function is 'Adaptive feasible', suggesting that the mutation rate is adjusted based on the progress of the algorithm and only feasible solutions are considered. The crossover method is 'Scattered', which means genes are randomly taken from each parent during reproduction.

The tolerance function is set to 1e-6. This is the stopping criteria for the algorithm, indicating that if the difference in fitness between the best and worst individual is less than this value, the algorithm will stop. These parameters are crucial as they can greatly affect the performance of the GA and are typically chosen based on the specific problem and through a process of trial and error.

The AM2HN model prediction accuracy was checked through a comparison of the model response with the ADM1 data. The search space is defined by the upper range Ub and lower range Lb of the seven parameters $[k_1, \mu_{1\max}, k_{S1}, \mu_{2\max}, k_{S1}, k_i, k_{hy}]$ as follows:

$$Lb = [5, 0.1, 0.1, 0.1, 1, 20, 1],$$

$$Ub = [150, 5, 8, 5, 10, 300, 10]$$

And the objective function is chosen as the logarithmic sum absolute of errors:

$$Obj = \ln(X_1) - \ln(X_{1m}) + \ln(X_2) - \ln(X_{2m}) + \ln(S_1) - \ln(S_{1m}) \quad (9)$$

B) Multiple linear regression

On the other hand, the 'regress' function in Matlab is a powerful tool for performing MLR. It estimates coefficients for a model based on predictor variables and responses. The function can also provide confidence intervals for these estimates, residuals of the model, and intervals for outlier detection. Additionally, it can return statistics such as the R squared statistic, F-statistic, p-value, and an estimate of the error variance. This function is particularly useful for iterative model fitting in a loop.

$$\beta = \min(S) = \min \sum_{i=1}^m [y_i - (\beta_i x_i + \epsilon)]^2 \quad (10)$$

Where β the regression coefficients and S are is the least squares criterion, x is the observation vector, ϵ is the error margin and m is the total number of observations.

The matrix formula is:

$$\min \|Y - X\beta\|^2 \quad (11)$$

Hence, the regression coefficients are calculated as follows:

$$\beta = (X'X)^{-1} X'Y \quad (12)$$

Where $(X'X)^{-1}$ must be invertible.

The p-values reported in tables 4 and 5, were obtained by using the function linhyptest on Matlab. This function requires the following inputs: the estimated parameters and the covariance matrix; the value of the null hypothesis (zero or other specific value); and the degrees of freedom of the covariance matrix.

Table 2. Residuals of the AM2HN using the estimated parameters via MLR with various HRT values

HRT	S1	X1	X2
5	0.3	-1.9	1.1
10	0.1	-2	1.2
15	5.0e-02	-2	1.1
20	7.0e-02	-2.2	1.1
25	0	-2.2	1
30	0	-2.3	0.9
35	0	-2.5	0.8

Table 3. Residuals of the AM2HN using the estimated parameters via Genetic algorithm with various HRT values

HRT	S1	X1	X2
5	0.3445	0.4284	1.1087
10	0.0966	0.3365	0.1633
15	0.0616	0.2502	0.0909
20	0.0579	0.167	0.0154
25	0.0521	0.0949	-0.0478
30	0.0411	0.023	-0.11
35	0.0372	-0.0289	-0.1616

Table 4. P-value relation to estimates obtained by GA

Parameter	μ_1 max + k_{s1}	k_1	μ_2 max + k_{s2} + k_f	$k_2 + k_3$	$k_4 + k_5$	k_6	k_{HVD}
P-value	0.02	4.3×10^{-05}	2.5×10^{-07}	0.2	4.5×10^{-05}	4.9×10^{-15}	4.6×10^{-07}

Table 5. p-value relation to estimates obtained by MLR

Parameter	μ_1 max + k_{s1}	k_1	μ_2 max + k_{s2} + k_f	$k_2 + k_3$	$k_4 + k_5$	k_6	k_{HVD}
P-value	1.6×10^{-2}	1.8×10^{-5}	3.10^{-8}	0.06	8×10^{-10}	4.5×10^{-15}	1.5×10^{-9}

The table 4 presents the p-values for the parameters estimated by Genetic Algorithm (GA). The p-value is a statistical measure that helps determine whether their hypotheses are correct. Typically, a smaller p-value means that the parameter is statistically significant.

In conclusion, the MLR method seems to provide a more statistically significant model for ADM1 data compared to the GA method, except for the parameter " $k_2 + k_3$ ". This can be explained by the fact that the function 'regress' in Matlab does not have constraints on the upper and lower bounds of the estimates (that results in the use of negative estimates in many situations which does not have any physical sense), unlike with GA.

However, it's important to note that statistical significance does not necessarily imply that the model is better at predicting or explaining the data as it is shown in [9]. The same estimates obtained by Chaabna et al. was tested in this study.

V. CONCLUSION

In conclusion, our study presented a comparative analysis of residuals from two parameter estimation techniques: Multiple Linear Regression (MLR) and Genetic Algorithms (GA).

The comparison of p-values obtained from both techniques shows that MLR provides more statistically significant results for most parameters. This suggests that MLR might be a more suitable method for modeling this particular dataset. However, it's important to remember that statistical significance does not necessarily equate to a better predictive or explanatory model since the function regress used in this study does not respect lower and upper bounds of the estimates which results negative estimates in many situations. Therefore, other model evaluation metrics should also be considered to determine the best modeling approach. It's also crucial to understand the underlying assumptions of each method and ensure that they are met by the data.

VI. ACKNOWLEDGMENTS

This work was supported by the PRFU project A01L08UN190120220002.

VII. REFERENCES

- [1] Noykova N, Müller TG, Gyllenberg M, Timmer J. Quantitative analyses of anaerobic wastewater treatment processes: identifiability and parameter

- estimation. *Biotechnology and bioengineering*. 2002 Apr 5;78(1):89-103.
- [2] Batstone DJ, Torrijos M, Ruiz C, Schmidt JE. Use of an anaerobic sequencing batch reactor for parameter estimation in modelling of anaerobic digestion. *Water Science and Technology*. 2004 Nov 1;50(10):295-303.
- [3] Fan Z, Ren Z, Chen A, Feng X, Wang W. A method for parameter identification of combined integrating systems. *Mathematical Problems in Engineering*. 2020 Feb 12;2020:1-3. J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [4] Wu RC, Tseng YW, Chen CY. Estimating parameters of the induction machine by the polynomial regression. *Applied Sciences*. 2018 Jul 1;8(7):1073.
- [5] Iweka SC, Owuama KC, Chukwunke JL, Falowo OA. Optimization of biogas yield from anaerobic co-digestion of corn-chaff and cow dung digestate: RSM and python approach. *Heliyon*. 2021 Nov 1;7(11).
- [6] Liu Y, Gunawan R. Bioprocess optimization under uncertainty using ensemble modeling. *Journal of biotechnology*. 2017 Feb 20;244:34-44
- [7] Tauš P, Kudelas D, Taušová M, Gabániová E. Statistical Approach for Assessing the Suitability of Substrates for a Biogas Plant. *Sustainability*. 2020 Oct 30;12(21):9044.
- [8] Hassam S, Ficara E, Leva A, Harmand J. A generic and systematic procedure to derive a simplified model from the anaerobic digestion model No. 1 (ADM1). *Biochemical Engineering Journal*. 2015 Jul 15;99:193-203.
- [9] Chaabna A, Semcheddine S. Genetic algorithm based identification of biogas production model from wastewater via anaerobic digestion model no. 1. *International Journal of Information Technology*. 2023 Mar;15(3):1465-72.
- [10] Silva I, Jorge C, Brito L, Duarte E. A pig slurry feast / famine feeding regime strategy to improve mesophilic anaerobic digestion efficiency and digestate hygienisation. *Waste Management & Research*. 2021 Jul;39(7):947-55.
- [11] Goldberg DE (1989) *Genetic algorithms in search, optimization and machine learning*, 1st edn. Addison-Wesley Longman Publishing Co., Inc
- [12] Michalewicz Z (1994) *Genetic algorithms + data structures = evolution programs*, 2nd extended edn. Springer-Verlag