CHEMICAL ENGINEERING TRANSACTIONS
VOL. 119, 2025
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: David Bogle, Flavio Manenti, Piero Salatino

ISBN 979-12-81206-20-5; **ISSN** 2283-9216
DOI: 10.3303/CET25119070

Advanced Steady-State Detection in Biogas Plant Data Using Statistical and Machine Learning Techniques

Loretta Salano, Flavio Manenti,Mattia Vallerio*

Politecnico di Milano, Dipartimento di Chimica, Materiali ed Ing. Chimica "G.Natta", P.zza Leonardo da Vinci 32, 20133, Milano.Italia
mattia.vallerio@polimi.it

The application of biogas in the industrial energy transition has strongly emerged as a relevant feedstock for chemicals and energy production, especially in the context of biogas reforming. Effective steady-state detection is crucial for optimizing the production processes, where stable operation directly impacts hydrogen yield and system efficiency. This study tests three steady-state detection (SSD) techniques on two datasets from a demo-scale biogas plant: (i) a statistical hypothesis testing approach, (ii) a trend-based sliding window method, and (iii) the machine learning-based isolation forest (IF) algorithm. A parameter sensitivity analysis was performed to optimize each technique's performance on the biogas plant data. Results indicate that the statistical methods are strongly influenced by the operative parameters, while the IF detects simultaneously outliers and transient values. The integration of an IF algorithm in the detection framework is suggested to enhance reliability in real-time monitoring of biogas reforming processes. This study shows the potential of advanced SSD methods in analysing biogas plant operations to improve hydrogen production and optimal control. Future research will focus on developing a unified detection framework and refining machine-learning models for real-time implementation.

1. Introduction

The EU's sustainability goals for 2050 are pushing the chemical industry towards a substantial transition (European Commission, no date). Amongst the renewable energy sources, biogas is valuable for energy and chemical production. Only in Italy, over a thousand plants are currently installed for energy production. Still, many projects for biomethane and chemical production applications have emerged, thanks also to the PNRR-funded projects (Gazzetta Ufficiale Della Repubblica Italiana, 2022).
In the process industry, data analysis is very important for optimal plant operation. This discipline is fundamental for the extrapolation of accurate plant models, especially for the development of new process layouts. Since most industrial plants aim at optimal nominal productivity, steady-state operation detection is fundamental. If a correct model is applied at the wrong time, types I and II errors (false positives, false negatives), biased or inaccurate parameter estimates, and ultimately inappropriate decisions could be made on how to move the system. Serious violation of the steady-state assumption may result in unstable operation when online, real-time or closed-loop optimization is applied (Kelly and Hedengren, 2013).
Mhamdi et al. summarised the application of statistical tests over time windows for the online optimization of a multi-stage flash desalination plant. They (a) performed a Student t-test on a linear regressed slope over the time window, (b) a Student t-test on two recently computed means with pooled standard deviations from two adjacent windows and (c) finally, an F-test on two recently computed standard deviations either from two adjacent windows or the same window but using two different filtered means (Mhamdi et al., 2010). Most industrial implementations of SSD use a form of (b), known as the mathematical theory of evidence (Narasimhan, Kao and Mah, 1987). Kelly and Hedengren (2013)provided a modified model that corrects the mean and standard deviation for the drift component to obtain more accurate results. Further on, Dalheim et al. (2020) further implemented a modified window-slided algorithm to detect steady-state intervals along. Tao et al. proposed an improved method for steady-state identification, applying an adaptive polynomial filtering steady-state identification (Tao et al., 2012). In recent years, machine learning has paved its way into data analysis due

Paper Received: 25 May 2025; Revised: 13 July 2025; Accepted: 4 October 2025
Please cite this article as: Salano L., Manenti F., Vallerio M., 2025, Advanced Steady-state Detection in Biogas Plant Data Using Statistical and Machine Learning Techniques, Chemical Engineering Transactions, 119, 415-420 DOI:10.3303/CET25119070

to its versatility. Isolation forest-based algorithms have gained popularity in outliers' detection and steady-state detection due to their capability to handle high-dimensional data and non-linear relationships effectively (Heigl et al., 2021).
The current work compares two statistical methods taken from literature and a machine learning approach for the detection of steady-state intervals on a set of industrial data derived from the operation of a biogas reforming plant.

2. Methodology

The industrial data tested in the current work was obtained from a demonstrative biogas plant campaign (Salano et al., 2025). The collected measurements regard the mass and volumetric flow rate between inputs and outputs of the system. These are fundamental for data reconciliation to ensure the mass balance is respected and the unit operations' models are validated. A block scheme of the plant is reported in Figure 1.

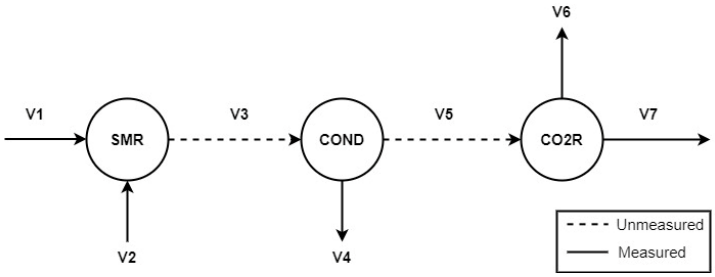

Figure 1 - Block scheme of the biogas plant. V1 is the biogas entering the steam methane reforming (SMR) reactor and water (V2). The product is sent to a condenser (COND) for water removal (V4) from the product. The syngas is upgraded to a lower tenor of CO2 through a water column (CO2R), after which the final product is recovered (V7).

The dataset also shows the dynamics phases for the startup and shutdown of the reformer unit. The endothermic catalytic process requires ramp phases for heating and cooling the unit. Dynamics of these phases are available and will be used to test the reliability of the steady state detection (SSD) algorithms. Further on, the product streams are analyzed to detect nominal production conditions. Two process variables are studied for each methodology: biogas fed to the reforming reactor (V1) and the product flow rate (V7). The criteria applied for steady-state detection is the intrinsic error of the instrument, of which values are reported in Table 1.

Table 1 - Process measurement device intrinsic error

Values	Measurement Device	Intrinsic error
V1	Coriolis Flowmeter	0.75%
Temperature	Thermocouple type K	0.4%

2.1 Statistical approach one (SSD1)

Kelly et al. proposed an SSD based on the assumption that the processed signal may be operating with a non-zero slope multiplied by its relative time within the window. The mathematical expression is:

$$x_t = mt + \mu + a_t \tag{1}$$

Where m is the deterministic drift component, μ is the mean of the hypothetical stationary process that will also equal the sample mean over the time window with zero slope, and a_t is the random error series. The drift component is obtained as the arithmetic average of $x_t - x_{t-1}$ with n sampled values of x_t.

$$\mu = \frac{1}{n}(\sum x_t - m\sum t) \tag{2}$$

$$\sigma_a = \sqrt{\frac{1}{n-2}\sum(x_t - mt - \mu)^2} \tag{3}$$

$$|x_t - \mu| \leq t_{crit} \cdot \sigma_a \tag{4}$$

The algorithm indicates the probability of each interval belonging to a steady state interval or transient one. A threshold value of 85% is requested to label the values as steady state. This approach requires minimal computational effort, as it requires only the evaluation of statistical parameters on a fixed window. The sampling time is chosen equal to three times the time-constant of the process divided by the sampling period, for the case under study, equal to 90 minutes while the critical value t_{crit} is equal to 2, as suggested by the referenced work.

2.2 Statistical slide-windows approach (SSD2)

Dalheim et al. (2020) tested a computationally efficient method on data from ship operation, particularly the engine power control system. The algorithm is based on the hypothesis that the signal can be modelled by a deterministic linear trend model, regardless of the window selection. The definition is the same as in Eq. (1), but m is calculated by ordinary least squares estimation (Eq. (5)).

$$m = \frac{\sum t\, x_t - \frac{1}{n}\sum t \sum x_t}{\sum t^2 - \frac{1}{n}(\sum t)^2} \tag{5}$$

$$\sigma_m = \frac{\sigma_a}{\sqrt{(t-\bar{t})^2}} \tag{6}$$

$$t_1 = \frac{m}{\sigma_m} \tag{7}$$

To determine whether the process is in a steady state, a statistical hypothesis test is performed on m. The null hypothesis H_0 assumes that the signal is stationary within the window, meaning that $m = 0$, there is no significant linear trend in the data. If the null hypothesis holds, the process is in steady state. Conversely, rejecting H_0 (i.e., finding $|t_1| > t_{\frac{\alpha}{2},n-2}$) implies that a significant trend exists in the data, indicating non-stationary behavior. The steady-state detector has two parameters: (i) significance level α, representing the probability of a type I error and rejecting the null hypothesis, and (ii) the window length n, defined in the number of samples. The tuning of said parameters is fundamental for the algorithm's performance and will be discussed in the results section.

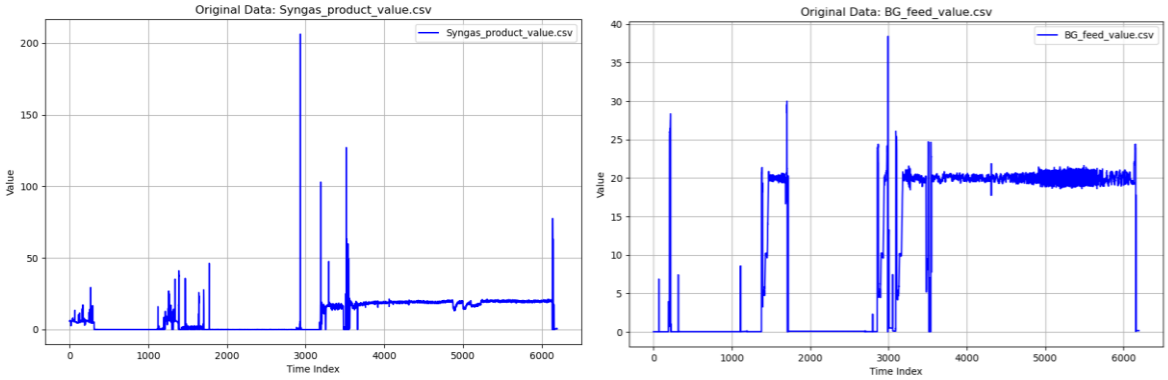

Figure 2- Raw data deived from the demonstative biogas plant: on the left the synags production flowrate (V7) while on the right the biogas feed flowrate (BG).

2.3 Machine Learning approach (IF)

Isolation forest (IF) shows great flexibility in the recognition of outliers in data sets. In previous applications, this approach has performed best for outliers' detection (Bouman et al. 2023, Salano et al. 2025). The working principle rests on the idea of isolating data points by randomly selecting a feature and then randomly choosing a split value between the minimum and maximum values of the selected feature. The random state controls the pseudo-randomness of the selection, for this study, it is kept equal to 42 from previous testing (Salano *et al.*, 2024). The contamination factor is proportional to the outliers expected in the dataset and should be tuned following the data behavior.

Table 2 - Sensitivity results for the biogas feed stream, furnace temperature and product stream.

Methodology	Parameter			Steady States
SSD1	t_{crit}		2	4770
SSD2	$n = 10$	α	0.01	5260
			0.1	4499
			0.5	3465
			0.8	2926
	$n = 50$		0.01	5606
			0.1	5532
			0.5	5382
			0.8	5327
IF	ϵ		0.1	5577
			0.2	4957
			0.4	3718
			0.5	3100
Methodology	Parameter			Steady States
SSD1	t_{crit}		2	5296
SSD2	$n = 10$	α	0.01	5972
			0.1	5408
			0.5	3916
			0.8	2971
	$n = 50$		0.01	5667
			0.1	5381
			0.5	4975
			0.8	4829
IF	ϵ		0.1	5576
			0.2	4959
			0.4	3720
			0.5	3107

3. Results

The raw collected data is reported in Figure 2, the first half of the time index series is associated with startup routines and intermittent shutdowns of the system. The nature of the data provides appropriate training for the method's ability to detect steady-state intervals.

The main results are reported in Table 2. The statistical methods perform well but tend to recognize outliers as steady-state points if not tuned optimally. SSD1, as expected, shows low flexibility, as the investigation is carried out along a fixed time span rather than a rolling window. The choice of the critical parameter is also fixed from the taken reference. Appropriate tuning might change the overall performance, but it is not an "a priori" information. SSD2, instead, allows for the choice of a rolling window time span, this allows to detect shifts in the data more robustly. For higher values of n the higher steady state points are detected, while the opposite trend is reported for α values. It is important to highlight that in this study outliers are not removed from the data, but rather included to highlight the potential of the machine learning approach.

In SSD2, higher steady-state points tend to include random outliers, as observable in Figure 3. It is interesting to compare the two statistical methods, SSD1 performs well with the literature parameters, for example around time step 5000 it doesn't blindly band all the changing points as steady state, like SSD2 in Figure 3 for example. The IF provides robust results, allowing the simultaneous identification of outliers and steady-state phases. Considering time step 5000 for both measured values, it is the only approach that consider noise values and identifies a more robust steady state for reconciliation purposes. The variation in the contamination factor impacts the total number of steady-state points. Even for a high contamination value, the identified steady state points don't overlap outliers. The method differentiates the two families and provides a simultaneous identification. It is interesting to observe that SSD2 is the only method that identifies some relevant steady-state data between time index 2800 and 3500, where the startup ramps are carried out for the SMR unit. The lack of identification from the IF algorithm is interconnected with the nature of the dataset. The authors have found that contamination values lower than 0.1 might be beneficial when combined with the specification of the number of

estimators in the algorithm. Figure 4 reports, in fact, results for a low contamination value of 0.07, a random state of 42 and a fixed number of estimators of 300 and for a higher contamination of 0.2 and random state of 500. Results show a higher sensitivity to detection for the first case, but also more false steady states.
Further on, a normalization of the data is helpful for this application, as well as considering the rate of change as the training data set, rather than the raw data. Results are reported in Figure 4. The performance of the IF

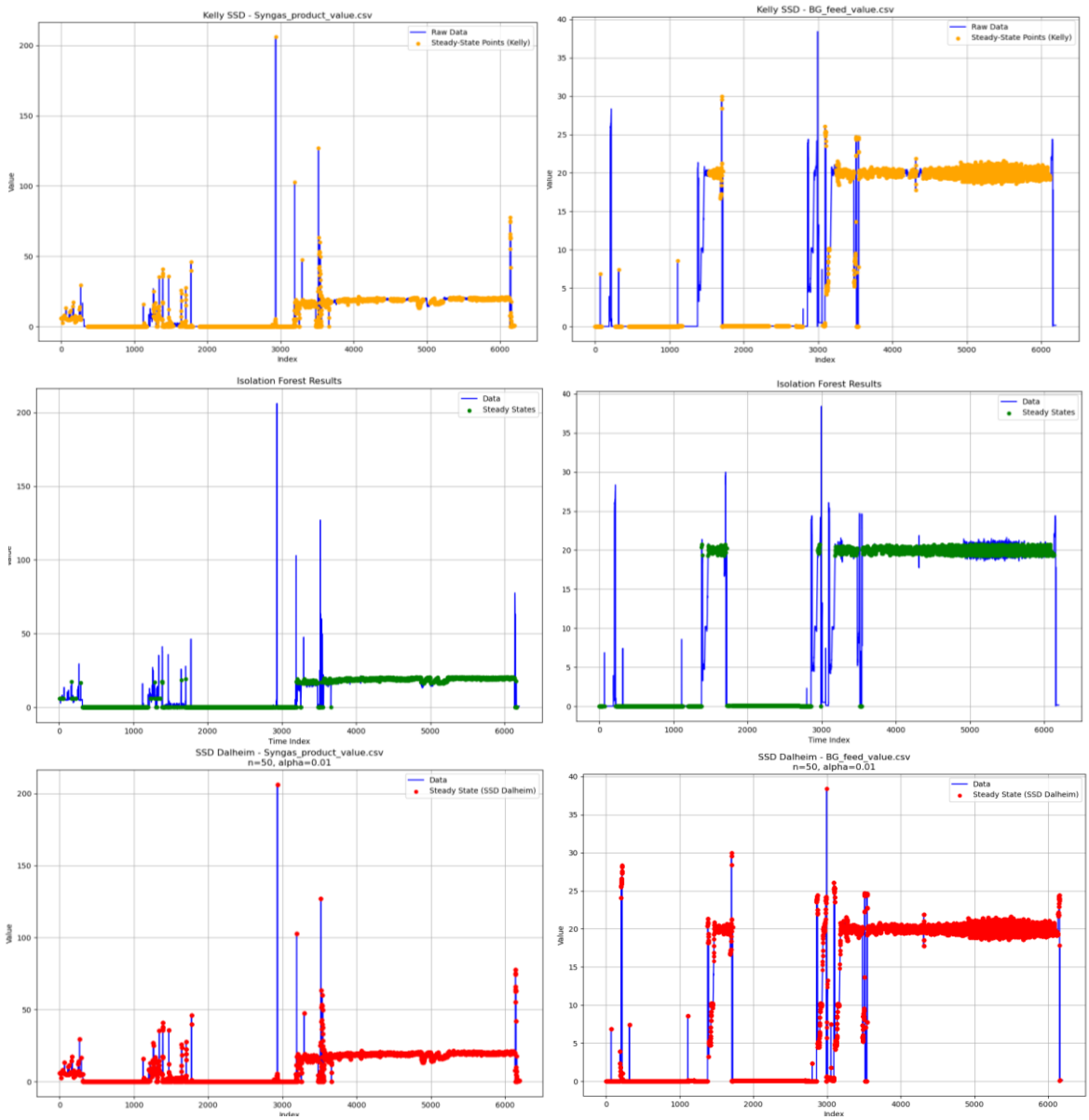

Figure 3 - Steady state detection. Top: SSD1 with sample time of 90 minutes. Middle: IF contamination of 0.1 and random state 42. Bottom: SSD2 with alpha of 0.01 and sample time of 50 minutes.

isn't strongly affected by the training set, but rather by the finer refinement of the contamination parameter. Further on, given the operational time expected for the operation of the plant, the better performance of the algorithm across the complete time range outcomes the statistical ones.

4. Conclusions

This study compared three steady-state detection (SSD) techniques—two statistical methods and a machine learning approach—on the industrial data from a biogas reforming plant. The results highlight the strengths and limitations of each technique in detecting steady-state operation. The statistical hypothesis testing method (SSD1) demonstrates good interval detections but lacks flexibility on the dataset outliers, as it relies on fixed-

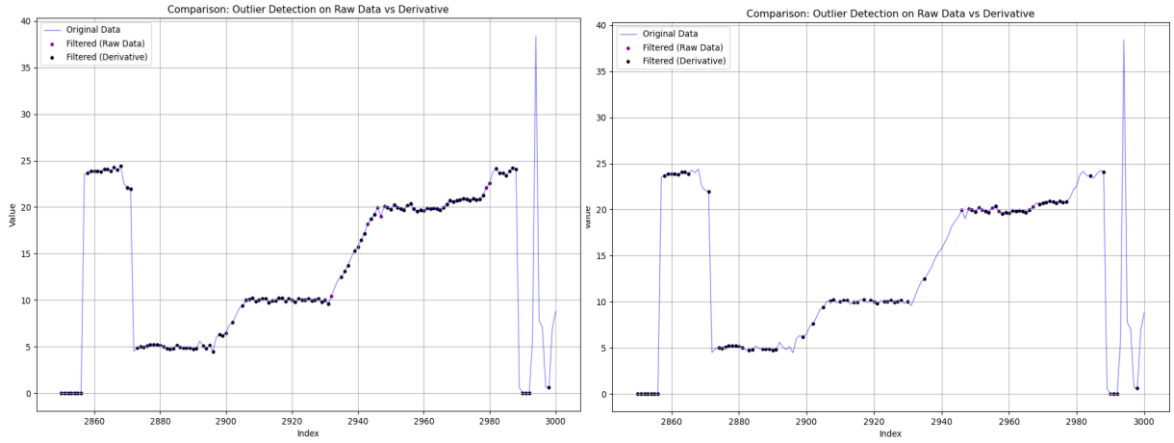

Figure 4 - Analysed data with IF training based on the raw data and their derivative along the startup interval: on the left for random state = 42, contamination = 0.07 and number of estimators = 300, on the right for random state = 500, contamination = 0.2 and number of estimators = 300.

window parameters. The trend-based sliding window approach (SSD2) improves adaptability by allowing tunable window lengths, though it remains sensitive to outliers when not optimally configured. The isolation forest (IF) algorithm, on the other hand, shows strong robust-ness in distinguishing steady-state conditions from transient phases and outliers, making it a promising candidate for real-time monitoring applications. The findings suggest that a hybrid approach, integrating statistical and machine learning methods, could enhance steady-state detection accuracy while maintaining computational efficiency. Future work will focus on refining parameter selection strategies, integrating domain-specific constraints, and testing the hybrid detection framework in a real-time industrial setting. Additionally, further development of adaptive machine learning techniques could improve generalizability across different process conditions, advancing steady-state detection capabilities in biogas and broader industrial applications.

References

Bouman, R., Bukhsh, Z. and Heskes, T. (2023) 'Unsupervised anomaly detection algorithms on real-world data: how many do we need?', 25, pp. 1–34. Available at: http://arxiv.org/abs/2305.00735.

Dalheim, Ø.Ø. and Steen, S. (2020) 'A computationally efficient method for identification of steady state in time series data from ship monitoring', *Journal of Ocean Engineering and Science*, 5(4), pp. 333–345. Available at: https://doi.org/10.1016/J.JOES.2020.01.003.

European Commission (no date) *EU 2050 long-term strategy*.

Gazzetta Ufficiale Della Repubblica Italiana (2022) 'Decreto Ministeriale 15 settembre 2022'.

Heigl, M. *et al.* (2021) 'On the improvement of the isolation forest algorithm for outlier detection with streaming data', *Electronics (Switzerland)*, 10(13), pp. 1–26. Available at: https://doi.org/10.3390/electronics10131534.

Kelly, J.D. and Hedengren, J.D. (2013) 'A steady-state detection (SSD) algorithm to detect non-stationary drifts in processes', *Journal of Process Control*, 23(3), pp. 326–331. Available at: https://doi.org/10.1016/J.JPROCONT.2012.12.001.

Mhamdi, A. *et al.* (2010) 'On-line optimization of MSF desalination plants', *Thermal Desalination processes*, I.

Narasimhan, S., Kao, C.S. and Mah, R.S.H. (1987) 'Detecting changes of steady states using the mathematical theory of evidence', *AIChE Journal*, 33(11), pp. 1930–1932. Available at: https://doi.org/10.1002/aic.690331125.

Salano, L., Vallerio, M. and Moioli, E. (2025) 'Industrial scale biogas reforming modelling and validation', *Chemical Engineering Journal*, 510(March), p. 160871. Available at: https://doi.org/10.1016/j.cej.2025.160871.

Tao, L. *et al.* (2012) 'Steady-state identification with gross errors for industrial process units', *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*, pp. 4151–4154. Available at: https://doi.org/10.1109/WCICA.2012.6359172.