BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 74(4), 2026, Article number: e158306
DOI: 10.24425/bpasts.2026.158306



# Model goodness of fit evaluation based on a fuzzy inference system in virtual commissioning

**Łukasz GLODEK**[1]*, **Anna GLODEK**[2], **Witold NOCOŃ**[2], **and Szymon BYSKO**[1]

[1] PROPOINT S.A., Gliwice, Poland

[2] Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland

**Abstract.** Modern industrial plants are becoming increasingly complex, resulting in the need for rapid testing and validation of industrial automation systems. To meet the requirements mentioned above, new simulation techniques, like virtual commissioning (VC), can be employed, as they allow for identifying process bottlenecks at the very beginning of the commissioning process. Moreover, it has also been used for maintenance operator training. The essential stage of VC is verification of the model of a commissioned plant quality – model goodness of fit. A plethora of measures are used for model goodness of fit evaluation, but each is characterized by a different range of values and interpretations. Thus, the best idea is to use the hybrid approach for model goodness of fit evaluation, combining the information from different measures. In order to create a flexible system for decision-making, if a model quality is good and sufficient to be used in VC, the Virtual-Commissioning-Model Fuzzy Coefficient (VCMF) is introduced based on the Takagi-Sugeno-Kang fuzzy-inference system. It considers knowledge of virtual commissioning of industrial automation systems and information carried by different methods of goodness of fit evaluation (NRMSE, ME, MAE, and MIA). VCMF was based on data from the belt conveyor, which was thoroughly analyzed. Current, velocity, and torque time series underwent the data pre-processing and analysis methods, which resulted in obtaining a model. VCMF allows for differentiating models into those that can be used in VC and those that cannot. The threshold value was defined by Gaussian Mixture Modeling and Bayesian Information Criterion.

**Keywords:** virtual commissioning; model goodness of fit; fuzzy logic; industrial automation systems.

## 1. INTRODUCTION

The complexity of modern industrial plants is continuously growing, which leads to the necessity of rapid testing and validation of industrial automation systems. The answer to this demand is virtual commissioning, one of the simulation techniques. It allows for analyzing the operation of various components and identifying bottlenecks in the entire production process at an early stage of the production line reconstruction. In addition, virtual commissioning (VC) is increasingly being used to train production line operators. One of the crucial stages of virtual commissioning is verifying the quality of the mathematical model of the commissioned object. A variety of metrics can be applied to assess model quality, each defined by its own value range and interpretive nuances. As a result, selecting the adequate measure to assess model quality is challenging, and in many cases the information carried by several measures would be needed to assess the overall quality of the model. A review of literature studies focused on model goodness of fit evaluation indicates that there is no single, unified and universal measure that can be used to assess the quality of a model. The following commonly used methods for goodness of fit evaluation were analyzed: Forecast Error (FE), Mean Forecast Error (MFE), Mean-Squared Error (MSE), Root Mean Squared Error (RMSE), Normalized Root Mean Squared Error (NRMSE), Root Relative Squared Error (RRSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (sMAPE), Median Absolute Error (MedAE), Mean Squared Logarithmic Error (MSLE), Explained Variance Score (EVS), Maximum Residual Error (ME), Coefficient of Variation (R2), Modified Index of Agreement (MIA), Relative Index of Agreement (RIA) and Pearson Product Moment Correlation Coefficient (PPMCC). Mean squared error and its modifications (MSE, RMSE), as well as mean error (MAE) and its percentage variant (MAPE) take values from the range of 0 to $+\infty$, which means that their single value does not carry too much information on regression performance. The coefficient of determination (R2) and sMAPE achieve high values only if the majority of the actual measurements have been appropriately planned. In [1] it has been proven that R2 is more informative and realistic than sMAPE and has no interpretative restrictions that MSE has, RMSE, MAE, and MAPE [1]. It is worth mentioning that after an analysis of the literature on virtual commissioning, the metrics that were used to assess the quality of the system models were RMSE (for appropriateness assessment) and R2 (for correctness assessment) [2]. However, compared to R2, MIA is a more reliable statistical measure, as it is more sensitive to differences between observed and predicted values. and variances, and is less sensitive to extreme values [3]. The MAE ignores small values and reflects only an error in predicting the largest value, which means that it cannot be used as a stand-alone measure of the assessment of model quality. In order to calculate

*e-mail: lukasz.glodek@propoint.pl

Manuscript submitted 2025-08-29, revised 2026-02-02, initially accepted for publication 2026-03-16, published in July 2026.

MFE, the assumption of the origin of the data must be met / time series from the same scale. ME allows us to eliminate models in which one peak appears with a height significantly higher than the other values, which could cause unjustified warnings or visual errors in HMI (Human Machine Interface) for a working virtual commissioning workstation. Moreover, scientific articles from 2022 on the subject of virtual commissioning [4, 5] indicate that there is a need to determine the quality of the model. Although some scientific articles discuss model validation, they frequently conclude with the assertion that 'the model reproduces the system's behavior,' offering no deeper investigation into how particular errors may lead to incorrect signals, warnings, or operational decisions in the VC environment [6]. Recent work introduces systematic methodologies for developing VC architectures (for example, through the use of SPES), demonstrating the growing maturity of the field [7]. Nevertheless, comparably advanced methods for evaluating model quality within these architectures remain absent. This work highlights several gaps in the existing literature. Although numerous studies discuss virtual commissioning, they rarely provide a systematic comparison of model evaluation metrics tailored to its specific requirements. Most works rely on traditional measures such as RMSE or $R^2$, without addressing their limitations or examining how modeling errors influence operational outcomes within VC environments, including HMI behavior, warnings, or decision-making processes. To address these gaps, this work systematizes the knowledge on commonly used goodness-of-fit metrics, introduces the Virtual Commissioning-Model-Fuzzy (VCMF) coefficient based on a Takagi–Sugeno–Kang (TSK) fuzzy-inference system, and demonstrates how multi-criteria evaluation can more effectively capture the practical implications of model inaccuracies in virtual commissioning. VCMF allows for the aggregation of information carried by various measures and dynamic expansion, thanks to the possibility of using expert knowledge in the virtual commissioning of industrial automation systems. The proposed coefficient was tested on measurement data collected during the belt conveyor operation and on a tank system, which was subjected to the pre-processing process using various statistical data analysis methods, time series forecasting, and Gaussian Mixture Modeling (GMM) with the use of Bayesian Information Criterion (BIC). As a result, the VCMF coefficient assigns models to two groups: well-matched to the data and incorrectly assigned ones. The obtained results were analyzed and compared with the described measures commonly used to assess the quality of models. To sum up, recent publications on virtual commissioning highlight the growing need for reliable model quality assessment, yet most studies conclude with the general statement that "the model reproduces the system's behavior," without analyzing how specific modeling errors propagate into incorrect signals, warnings, or operational decisions within the VC environment. At the same time, new methodological advances — such as systematic VC architectures based on SPES—demonstrate the increasing maturity of the field. However, equally mature, integrated, and practically applicable methods for evaluating model quality within these architectures remain absent. This work addresses these gaps by providing both a novel methodological contribution and a practical industrial tool. First, it systematizes and critically compares commonly used goodness-of-fit metrics specifically in the context of virtual commissioning, highlighting their limitations and practical implications for industrial automation systems. Moreover, it introduces VCMF, a new multi-criteria model quality indicator based on a Takagi–Sugeno–Kang fuzzy inference system. Unlike traditional single-metric approaches, VCMF aggregates information from multiple measures and incorporates expert knowledge, enabling dynamic adaptation to different industrial scenarios. Additionally, it demonstrates the industrial relevance of VCMF by applying it to real measurement data from a belt conveyor system and a tank system, processed using statistical analysis, time-series forecasting, and Gaussian Mixture Modeling (GMM) with Bayesian Information Criterion (BIC). The proposed coefficient effectively distinguishes well-matched models from poorly performing ones, offering a more interpretable and operationally meaningful assessment than conventional metrics. By providing a unified, extensible, and practically oriented model evaluation framework, this work contributes both to the scientific development of virtual commissioning methodologies and to their industrial applicability. The results support more reliable VC workflows, reduce the risk of incorrect HMI behavior or false alarms, and enhance the robustness of automation system validation. This work is a part of doctoral dissertation, the analysis and research were copied from the dissertation [8]: Ł. Glodek, "Application of Fuzzy Systems for Model Quality Assessment in Virtual Commissioning of Industrial Automation Systems" (Zastosowanie systemów rozmytych do oceny jakości modeli na potrzeby wirtualnego rozruchu systemów automatyki przemysłowej), Silesian University of Technology, Poland, 2024.

## 2. MATERIALS AND METHODS

### 2.1. Belt conveyor

A belt conveyor shown in Fig. 1 consists of three modules. The single module comprises a motor, inverter, pulley and transport belts.

**Fig. 1.** Belt conveyor controlled by PLC Siemens S7-300

The belt conveyor was instrumented to measure motor current, supply voltage, and belt speed. Motor current was acquired using a nonintrusive Hall-effect current sensor mounted on the motor supply line. The supply voltage was measured in parallel at the motor terminals using voltage probes. Belt speed was obtained from an incremental encoder installed on the drive pulley shaft and converted to linear speed using the pulley diameter. All signals were synchronously recorded using a data acquisition system with a fixed sampling rate. Sensor calibration and basic filtering were applied prior to analysis. During the belt conveyor's work, 40 time-series data of current, velocity and torque were gathered. For each measurement series, a Grey-Box model was created. The expert assessed 20 models as well-matched to the data ('good' model) and 20 as not well-matched to the data ('bad' model). The models were classified as 'bad' by the expert because they exhibited a systematic error (bias) and consistently underestimated the values, failed to reproduce the correct dynamics (the response was too slow), and reacted too weakly to changes in the input signal.

The control plant can be represented as the state-space model based on [9]:

$$\begin{cases} \dot{x} = Ax + Bu, \\ y = Cx + Du \end{cases} \tag{1}$$

with:

$$x = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \omega_1 \\ \omega_2 \end{bmatrix}, \tag{2}$$

$$u = \begin{bmatrix} M_m \\ M_L \end{bmatrix}, \tag{3}$$

where
$M_m$ – motor torque,
$M_L$ – load torque,
$\varphi_1$ – angular position of the drive side,
$\varphi_2$ – angular position of the load side,
$\omega_1$ – angular velocity of the drive side,
$\omega_2$ – angular velocity of the load side,
$k_t$ – torsional stiffness,
$c_t$ – torsional dumpling,
$B_1$, $B_2$ – viscous friction with corresponding moments of inertia $J_1$, $J_2$
and the torque $M_t = k_t(\varphi_1 - \varphi_2) + c_t(\omega_1 - \omega_2)$.

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\dfrac{k_t}{J_1} & \dfrac{k_t}{J_1} & -\dfrac{c_t + B_1}{J_1} & \dfrac{c_t}{J_1} \\ \dfrac{k_t}{J_2} & -\dfrac{k_t}{J_2} & \dfrac{c_t}{J_2} & -\dfrac{c + t + B_2}{J_2} \end{bmatrix}, \tag{4}$$

$$B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \dfrac{1}{J_1} & 0 \\ 0 & -\dfrac{1}{J_2} \end{bmatrix}, \tag{5}$$

$$y = \begin{bmatrix} \omega_1 \\ \omega_2 \\ M_t \end{bmatrix}, \tag{6}$$

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ k_t & -k_t & c_t & -c_t \end{bmatrix}, \tag{7}$$

$$D = 0. \tag{8}$$

### 2.2. Data pre-processing

A pre-processing process based on statistical analysis and time-series forecasting was performed to assess the data quality, which led to model selection (Fig. 2).

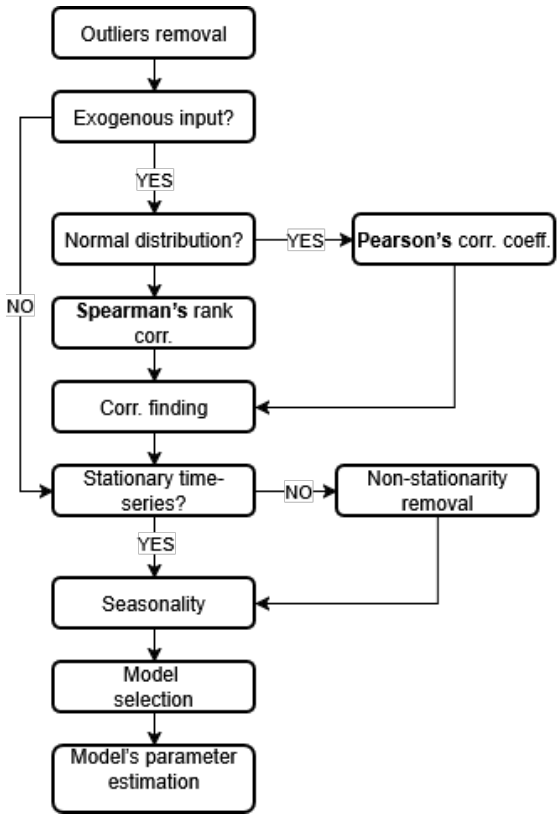


**Fig. 2.** Data pre-processing and analysis resulting in a model selection; *corr.* – correlation, *coeff.* – coefficient

A five-fold cross-validation procedure was employed to ensure a reliable performance assessment despite the limited dataset size. The full dataset was randomly partitioned into five equally sized subsets (folds). In each iteration, four folds were used for training and optimizing the fuzzy inference model, while the remaining fold served as an independent test set. This

process was repeated five times, so that every sample was used exactly once for testing. The final performance metrics were obtained by averaging the results across all five iterations. This approach provides a strict separation between training and testing data in every fold, prevents data leakage during model optimization, and yields a more robust estimate of the generalization capability of the model under small-sample conditions.

The pre-processing workflow consists of the following steps:

1. Outlier removal based on the box plot analysis
   Box plots for all the samples: current, velocity, and torque are in the same range, and the medians are similar (Fig. 3). The maximum current value equals the third quartile (Q3), which means that the average value between the median and the highest value equals the maximum. There are no outliers in the data [10].

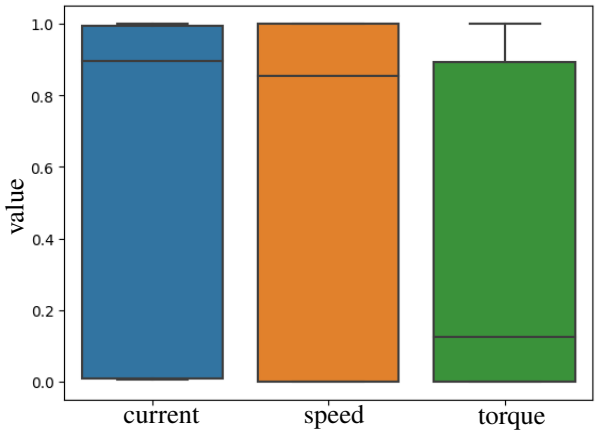


**Fig. 3.** Box plots for current, velocity and torque

2. Testing for normality: Shapiro-Wilk test
   Analyzing the results of the Shapiro-Wilk test (Table 1), it can be noticed that the data does not come from the normal distribution ($p$-value < 0.05). It implies the way of calculating the correlation between variables [11, 12].

**Table 1**
Results of normality test for current, velocity and torque data

| | $p$-value |
|---|---|
| *current* | $1.39\times10^{(}-28)$ |
| *speed* | $1.93\times10^{(}-28)$ |
| *torque* | $4.37\times10^{(}-29)$ |

3. Correlation between variables [13–15]
   Figure 4 shows Spearman's rank correlation between variables. After the heatmap analysis, it can be noticed that there is a strong relationship between the signals of current and torque.
   The correlation coefficient for current and velocity is equal to 0.90, which means a very strong correlation. The correlation between torque and current is also strong (0.83). The lowest correlation value is between current and torque

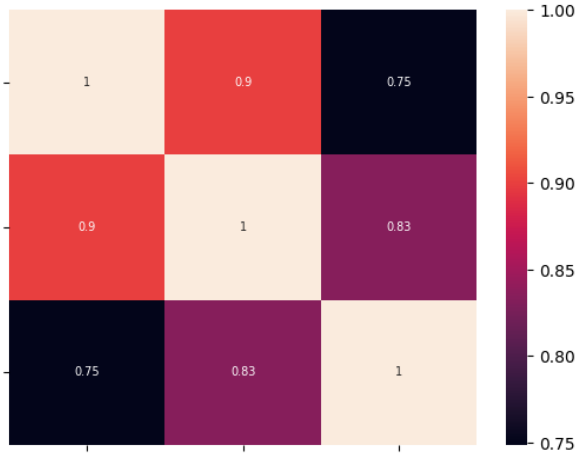


**Fig. 4.** Heatmap of the Spearman's rank correlation coefficient of current, velocity and torque

(0.75) [16]. Based on the values of Spearman's rank correlation coefficient, creating only two models is sufficient due to the high redundancy between the variables that represent current and velocity:
Current = f(velocity)
Torque = f(velocity)

4. The Augmented Dickey–Fuller Test for stationarity of time-series
   Analyzing the results of the Augmented Dickey–Fuller test for stationarity (Table 2), only the current time series is stationary ($p$-value < 0.05) [17].

**Table 2**
Results of the Augmented Dickey–Fuller test

| | Test statistic value | $p$-value |
|---|---|---|
| *current* | −3.03 | 0.03 |
| *velocity* | −2.82 | 0.06 |
| *torque* | −2.50 | 0.12 |

5. Seasonality
   Another crucial aspect is looking for a seasonal component in the data. After performing an additive decomposition, one can see a time series comprising three components: a remainder component, a seasonal component, and a trend-cycle component (Fig. 5). It can be noticed that the remainder part has a cyclic periods of significant amplitude. The trend is monotonically changing. Seasonality does not change over time, but its amplitude is constant over time. Thus, there is no need to remove the seasonal component from the time series data [18].
6. Model selection
   Autocorrelation function (ACF) for current exhibits decreasing tendency (Fig. 6).
   Analyzing the partial autocorrelation function (PACF) (Fig. 7), it can be noticed that there is a significant spike

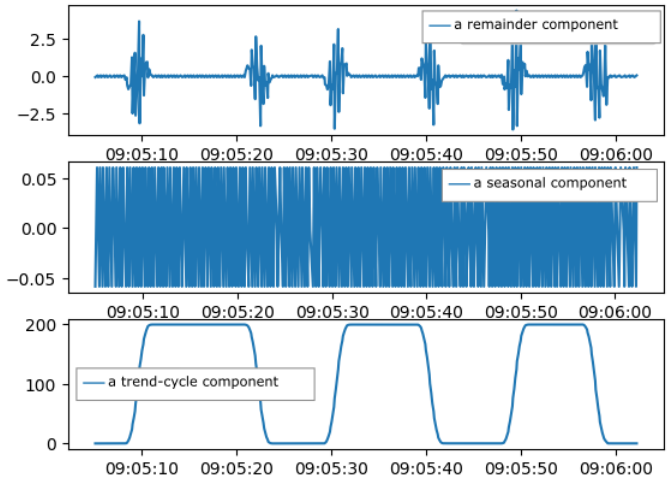


**Fig. 5.** Current time series decomposition into components: a remainder component, a seasonal component, a trend-cycle component

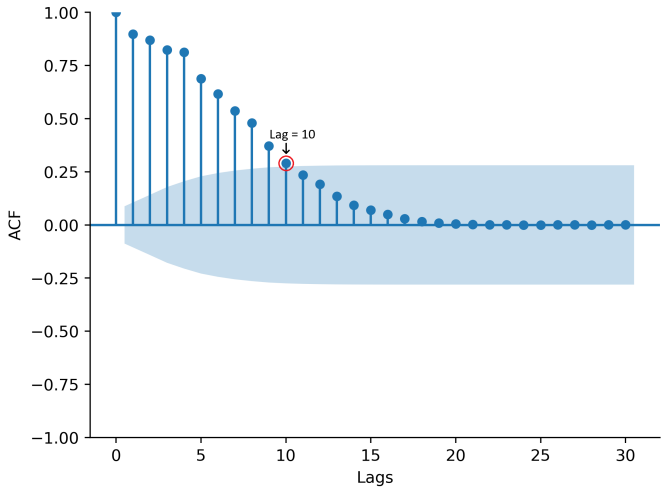


**Fig. 6.** Autocorrelation Function (ACF) for current with 5% significance limits for the autocorrelation

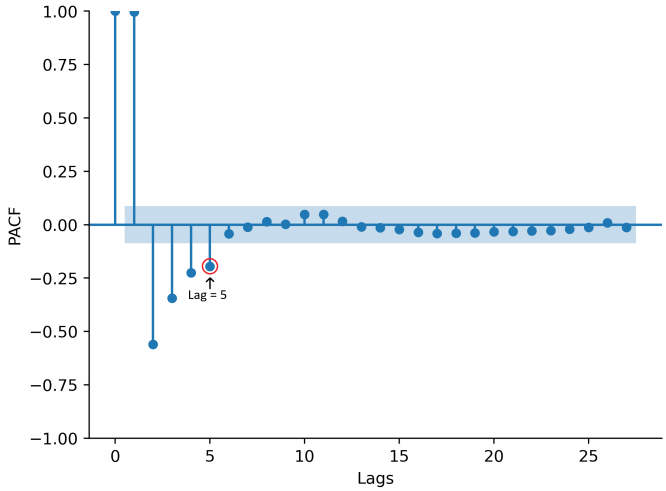


**Fig. 7.** Partial Autocorrelation Function (PACF) for current with 5% significance limits for the partial autocorrelation

at lag = 1, which decreases in subsequent lags, indicating a moving average in the data. Based on the number of lags outside the confidence interval (lag = 1 is autocorrelation), the polynomial degrees where defined: p = 5, q = 10 [19]. Analyzing the lag plots reveals nonlinearity in the data (Fig. 8). Taking into consideration the polynomial orders and the aforementioned outcomes, the NAR(5)MAX(10) model was selected. Parameters of the model were fit using the neural network – Multilayer Perceptron (Table 3). The neural network consists of the input layer, three hidden layers (each comprises 8, 16 and 16 neurons, respectively) and the output layer (1 neuron), 20 epochs. ReLU (Rectified Linear Unit), a widely used activation function in deep learning, was employed due to its ability to efficiently model nonlinear relationships [21]. For the torque, ACF, PAF, lag plots and parameter values were calculated in the similar way as for the current. At each iteration, the current model is compared with the best selected using the VCMF factor. If it is better, it becomes the new best model. Otherwise, it is discarded. The number of iterations is equal to the number of epochs.

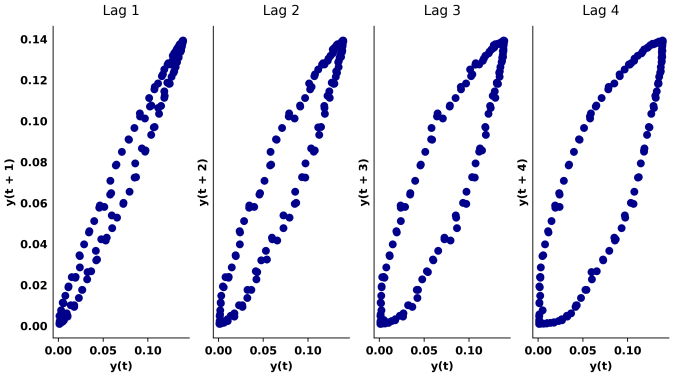


**Fig. 8.** Lag plots for current

**Table 3**
Parameters of NARMAX model estimated by the neural network

| $\theta_0$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|---|---|---|---|---|
| 0.2837 | 0.3603 | 0.5901 | 0.3791 | 0.3782 |
| $\theta_5$ | $\theta_6$ | $\theta_7$ | $\theta_8$ | $\theta_9$ |
| 0.4519 | 0.3274 | 0.1720 | 0.1812 | 0.0861 |
| $\phi_0$ | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\phi_4$ |
| 1.1886 | -0.1155 | -0.4539 | 0.3704 | -0.1294 |

In Fig. 9, a good fit between predicted current and torque values and the time series data can be observed. VCMF has identified 20 out of 20 models from the GOOD group as well-fitted to the data. Eighteen out of 20 models from the BAD group were identified correctly, and two models were wrongly identified.

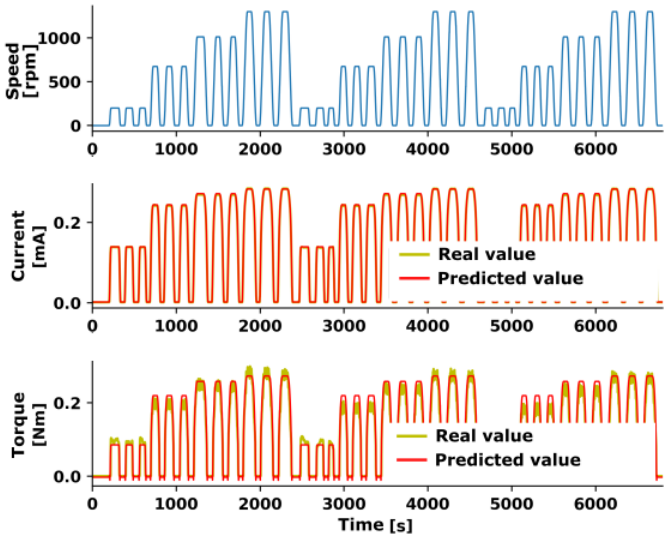


**Fig. 9.** Fit to a time series data using NAR(5)MAX(10) model for current and AR(3)X for torque

## 3. VIRTUAL COMMISSIONING-MODEL-FUZZY COEFFICIENT (VCMF)

Fuzzy logic allows for considering expert knowledge, often based on imprecise linguistic variables, which are not numerical (crisp) values. To incorporate increasingly evolving expert knowledge on virtual commissioning and modern automation systems control into a method for model goodness of fit evaluation, the VCMF coefficient based on the Takagi-Sugeno-Kang (TSK) fuzzy inference system, which combines information from four different well-known methods used for model goodness of fit evaluation: NRMSE, ME, MAE, and MIA, was introduced. The reason for considering only the methods mentioned above is that after analyzing estimated probability distribution functions, it can be noticed that some methods can differentiate the measurements into two groups: good and bad models (e.g., Fig. 10), as they show which range of values are likely to appear. For the remaining methods, like PPMCC or RIA, the

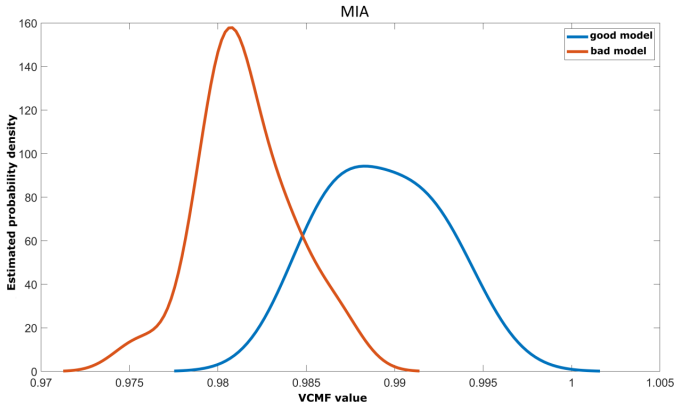


**Fig. 10.** Estimated density probability function (PDF) for MIA. The red line denotes the PDF for MIA value defining the "bad model", whereas the blue line denotes the PDF for MIA values identified as the "good model"

differences between good and bad models were not observed. Moreover, after analysis of different methods for model quality evaluation, it can be noticed that none of them is sufficient for assessing model quality for virtual commissioning purposes. What is more, many methods provide redundant information. It is worth mentioning that after reviewing the literature on virtual commissioning, the metrics that were taken to assess the quality of simulation models were RMSE (to assess appropriateness) and R2 (to assess correctness) [2]. However, compared to R2, the more reliable statistical measure is MIA, as it is more sensitive to differences between observed and predicted values of the averages and variances and is less sensitive to extreme values [3]. MAE ignores small values and reflects only the error in predicting the most significant value, which means it cannot be used as a stand-alone measure to assess the quality of the model [20, 21]. In order to calculate the MFE, it is necessary to meet the assumption of the origin of the data/time series from the same scale. ME allows the elimination of models in which a single peak significantly exceeds the other values, which could otherwise cause false alarms or errors in the HMI (Human–Machine Interface) visualization for the virtual commissioning workstation.

The TSK system was chosen because it is computationally efficient [22], which is essential from the industrial point of view. Modern automation systems are designed for complex physical plant systems and must manage the data from many sensors. The idea was presented in [23] and then it was extended.

The structure of created TSK fuzzy-inference system is shown in Fig. 11 [24]. The system's input values are numerical (NRMSE, ME, MAE, MIA). Therefore, the first phase of the process is fuzzification, where the input values are mapped into the values from the range <0; 1> based on the membership functions [24, 25].

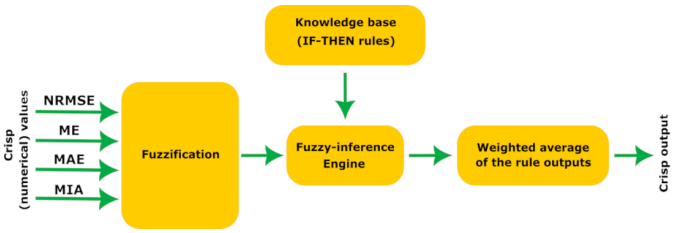


**Fig. 11.** A structure of the TSK fuzzy-inference system: 4 inputs (NRMSE, ME, MAE, MIA), 1 output (VCMF value)

Membership functions were defined based on experts' knowledge of virtual commissioning. The expert defined the permissible error and the shape of a membership function – Gaussian combination membership function [22, 24] (Fig. 12) (9). The reason for choosing Gaussian combination membership function is that its smooth and continuous nature allows for a probabilistic interpretation of membership grades, enabling reliable separation between well-matched (GOOD) and poorly matched (BAD) models. Furthermore, the gradual transitions provided by the Gaussian function improve robustness to noise and measurement uncertainty, which is essential for accurate

model validation and dependable virtual commissioning.

$$\mu_A(x;\sigma_1,c_1,\sigma_2,c_2)=\begin{cases} e^{-\frac{(x-c_1)^2}{2\sigma_1{}^2}}, & x\le c_1 \\ 1, & c_1<x\le c_2 \\ e^{-\frac{(x-c_2)^2}{2\sigma_2{}^2}}, & x>c_2 \end{cases} \quad (9)$$

$\sigma_1, \sigma_2$ – standard deviation,
$c_1, c_2$ – mean for each Gaussian function.

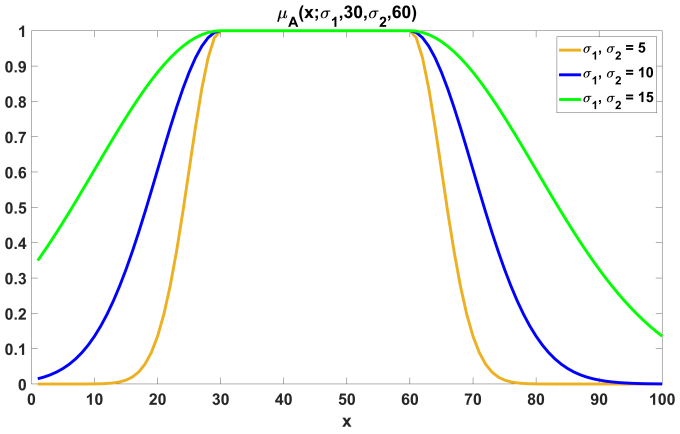


**Fig. 12.** Gaussian combination membership function for different $\sigma$ values

In order to assign the values (parameters), five methods were conducted: statistical (based on the intersection between the density functions of the two groups: good model and bad model), genetic algorithm (GA), particle swarm optimization (PSO), generalized pattern search (GPS), and simulated annealing algorithm (SA) [22].

The idea of selecting membership function parameters using optimization algorithms was presented in Fig. 13 [26]. Example: selection of membership function parameters using a genetic algorithm: membership functions are encoded in binary (as bit strings) [22]. The initial population consisted of 200 individuals. Individuals for the crossover were selected based on tournament selection. A crossover ratio of 0.8 resulted in 10 offspring

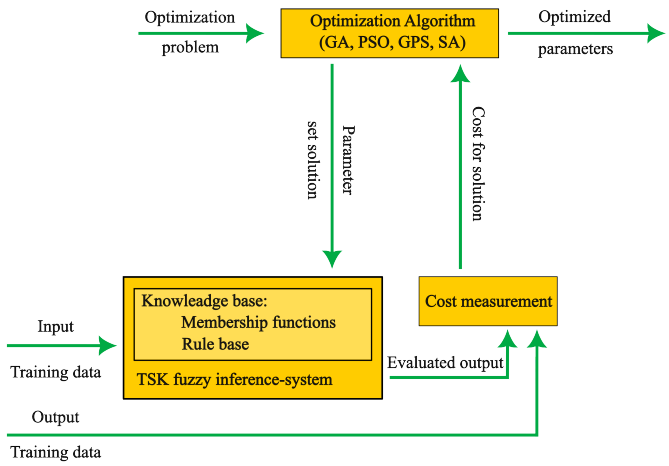


**Fig. 13.** A process of membership functions parameter selection [26]

individuals with the best fitness function values. The membership function values obtained using the genetic algorithm are presented in Table 4.

**Table 4**
GA parameter values of the membership function

| | $\sigma_1$ | $c_1$ | $\sigma_2$ | $c_2$ |
|---|---|---|---|---|
| NRMSE – GOOD | 0.05419 | 0.00476 | 0.00697 | 0.02412 |
| NRMSE – BAD | 0.00467 | 0.01990 | 0.00274 | 0.01010 |
| ME – GOOD | 0.31600 | −1.98000 | 4.14200 | 2.77800 |
| ME – BAD | 14.40000 | 35.85000 | 5.31000 | 107.50000 |
| MAE – GOOD | 0.00999 | 0.00098 | 0.00008 | 0.00281 |
| MAE – BAD | 0.00995 | 0.00135 | 0.00851 | 0.01090 |
| MIA – GOOD | 0.02675 | 0.04140 | 0.03332 | 0.54720 |
| MIA – BAD | 0.01920 | 0.75400 | 0.00002 | 0.98500 |

The PSO algorithm evaluates the objective function for each particle at every step and then determines its new velocity [22]. The convergence curve shows that the optimization cost reaches its minimum at approximately 40 iterations (Fig. 14). Parameters of the membership function optimized using PSO are presented in Table 5.

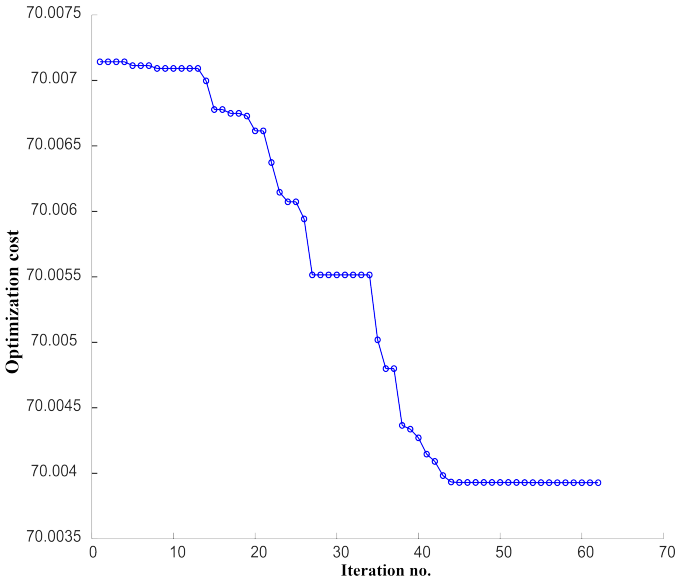


**Fig. 14.** Convergence plot of the optimization cost versus the iteration number (PSO)

In each iteration, simulated annealing samples a new candidate solution from a temperature-scaled probability distribution. Moves that reduce the objective are accepted unconditionally, while uphill moves are admitted with a temperature-dependent probability [27]. The cost function reaches its optimal value after 105 iterations (Fig. 15). Parameters of the membership function optimized using SA are presented in Table 6. The Generalized Search Algorithm searches for a sequence of points that

**Table 5**
Parameters of the membership function optimized using PSO

| | $\sigma_1$ | $c_1$ | $\sigma_2$ | $c_2$ |
|---|---|---|---|---|
| NRMSE – GOOD | 0.00756 | −0.00949 | 0.00423 | 0.00463 |
| NRMSE – BAD | 0.00446 | 0.02235 | 0.00329 | 0.10420 |
| ME – GOOD | 0.31600 | −1.98000 | 4.14200 | 2.77800 |
| ME – BAD | 14.40000 | 35.85000 | 5.31000 | 107.50000 |
| MAE – GOOD | 0.00142 | −0.00042 | 0.00142 | 0.00141 |
| MAE – BAD | 0.00995 | 0.00336 | 0.01000 | 0.01015 |
| MIA – GOOD | 0.00003 | 0.98470 | 0.92010 | 1.17400 |
| MIA – BAD | 0.98330 | 0.62230 | 1.00000 | 0.98210 |

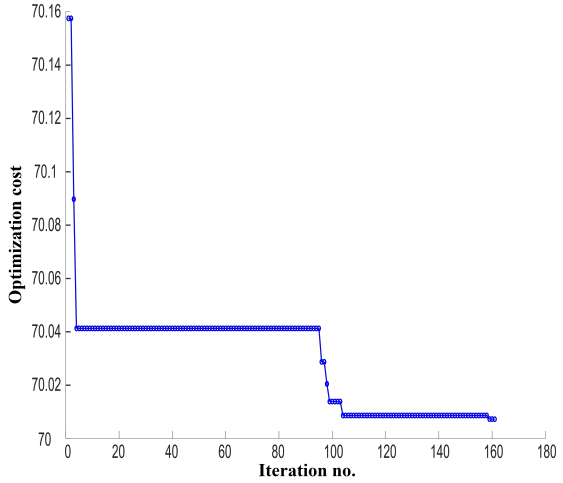


**Fig. 15.** Convergence plot of the optimization cost versus the iteration number (SA)

**Table 6**
Parameters of the membership function optimized using SA

| | $\sigma_1$ | $c_1$ | $\sigma_2$ | $c_2$ |
|---|---|---|---|---|
| NRMSE – GOOD | 0.07547 | −0.00001 | 0.08841 | 0.02236 |
| NRMSE – BAD | 0.02899 | 0.07821 | 0.07364 | 0.09027 |
| ME – GOOD | 0.31600 | −1.98000 | 4.14200 | 2.77800 |
| ME – BAD | 14.40000 | 35.85000 | 5.31000 | 107.50000 |
| MAE – GOOD | 0.00027 | 0.00047 | 0.00001 | 0.00289 |
| MAE – BAD | 0.00354 | 0.00845 | 0.00371 | 0.00340 |
| MIA – GOOD | 0.32530 | 0.80050 | 0.03354 | 0.27290 |
| MIA – BAD | 0.44720 | -0.13330 | 0.21420 | 0.97580 |

approach the optimal solution. After 26 iterations, it stopped because the mesh size ($9.5367 \times 10^{-7}$) became smaller than the mesh-size tolerance ($10^{-6}$) [22]. Parameters of the membership function optimized using GPS are presented in Table 7.

The values of the parameters obtained by all the methods were compared, and the Cohen's d effect size [28] analysis was performed to choose the best set of parameters. For the group

**Table 7**
Parameters of the membership function optimized using GPS

| | $\sigma_1$ | $c_1$ | $\sigma_2$ | $c_2$ |
|---|---|---|---|---|
| NRMSE – GOOD | 0.00756 | −0.00949 | 0.07454 | 0.06615 |
| NRMSE – BAD | 0.00531 | 0.09571 | 0.00329 | 0.10420 |
| ME – GOOD | 0.31600 | −1.98000 | 4.14200 | 2.77800 |
| ME – BAD | 14.40000 | 35.85000 | 5.31000 | 107.50000 |
| MAE – GOOD | 0.00142 | −0.00042 | 0.00001 | 0.00287 |
| MAE – BAD | 0.00001 | 0.00296 | 0.00004 | 0.01300 |
| MIA – GOOD | 0.98430 | 0.04128 | 0.07950 | 0.22450 |
| MIA – BAD | 0.06440 | −0.27470 | 0.00138 | 0.98520 |

GOOD the size effect between the data and ideal values was calculated according to the formula (10) [28]:

$$d_{GOOD} = \frac{\mu_{DI} - \mu_D}{s_{Dpooled}}, \tag{10}$$

where
$\mu_{DI}, \mu_D$ are the means from groups *GOOD IDEAL* and *GOOD*,
$s_{Dpooled}$ – pooled standard deviation.

The pooled standard deviation provides a single measure of spread for two or more independent groups by assuming they share a common population variance. It is computed as a weighted combination of the groups' individual variances, so samples with more observations have a greater influence on the final estimate [29]. For the BAD group the size effect value was calculated in a similar way. Analyzing the effect size values (Table 8), one can observe the difference between the means of the groups. The smaller the difference between the groups (GOOD and GOOD IDEAL, BAD and BAD IDEAL) the closer the groups are.

**Table 8**
Cohen's d effect size values

| Effect size | Good | Bad | Mean |
|---|---|---|---|
| *statistical method* | 1.33490 | 1.09110 | 1.21300 |
| *GA* | 0.87887 | 0.81763 | 0.84825 |
| *PSO* | 2.74348 | 0.53995 | 1.64172 |
| *GPS* | 291.58893 | 0.45033 | 146.01963 |
| *SA* | 10.28767 | 0.45033 | 5.36900 |

Based on those results, parameters assigned by GA were used to define the membership function, as they are characterized by the lowest mean value of the effect size.

Experts often use imprecise expressions, such as "The ME value should be quite low and simultaneously not exceed a certain value in order not to result in an unjustified exceedance of an alarm threshold." Due to that the knowledge base based on IF-THEN rules should be created. The knowledge base of the

TSK system is given by the following formula (11).

$$\mathscr{R} = \left\{\mathscr{R}\right\}_{i=1}^{I} = \left\{\text{if } \bigwedge_{n=1}^{N} x_{0n} \text{ is } A_n^{(i)} \text{ , then } y = f_i(\mathbf{x_0})\right\}_{i=1}^{I} \tag{11}$$

where $I$ – no. of rules in the knowledge base, $x_{0n}$ – input singleton,

$$\mathbf{x}_0 = [x_{01}, x_{02}, \ldots, x_{0n}]^T, \tag{12}$$

$A_n^{(i)}$ linguistic value of the premise of the rule,
$y = f_i(\mathbf{x_0})$, function in the $i$-th *IF-THEN* rule conclusion.

The system's knowledge base consists of the following six IF-THEN rules, which were constructed based on expert knowledge of virtual commissioning and industrial automation systems and assumptions in business contracts. The subtractive clustering method was used to generate the TSK fuzzy rules. All inputs were normalized into the $\langle 0, 1\rangle$ range. To obtain approximately six fuzzy rules with high prediction accuracy, the cluster radius was tuned in the range from 0.30 to 0.60 with a step of 0.15. The optimal value was found around $r = 0.38–0.40$, providing a good balance between model complexity and generalization capability for the dataset.

1. IF ME is BAD, THEN the QUALITY is BAD
2. IF NRMSE is BAD, THEN the QUALITY is BAD
3. IF NRMSE is BAD AND ME is BAD and MAE is BAD and MIA is BAD, THEN the QUALITY is BAD
4. IF NRMSE is GOOD and ME is GOOD and MAE is GOOD AND MIA is GOOD, THEN the QUALITY is GOOD
5. IF NRMSE is GOOD and ME is GOOD and MIA is GOOD, THEN the QUALITY is GOOD
6. IF NRMSE is GOOD and ME is GOOD and MAE is GOOD, THEN the QUALITY is GOOD

Figure 16 presents how the system works.

The system's final output is the weighted average of all rule outputs (13). The obtained value is multiplied by 100, which is the final VCMF value.

$$\text{Final output of the system} = \frac{\sum_{i=1}^{N} w_i z_i}{\sum_{i=1}^{N} w_i}. \tag{13}$$

In order to define the threshold over which the VCMF values are considered as a good model (that fits the data well) and below which – the model does not fit the data well (bad model), the distribution of VCMF values was estimated by undergoing Gaussian Mixture Modeling algorithm (GMM) [30] with three components characterized by normal probability distribution (Fig. 17). The number of components was defined by the lowest value of Bayesian Information Criterion (BIC) [31].

Finally, the threshold value for the data from the belt conveyor is 64.6.

VCMF correctly assigned 20 out of 20 models from the group GOOD as those well-matched to the data. However 18 out of 20 models from the group BAD were correctly identified, but 2 of them were incorrectly classified.

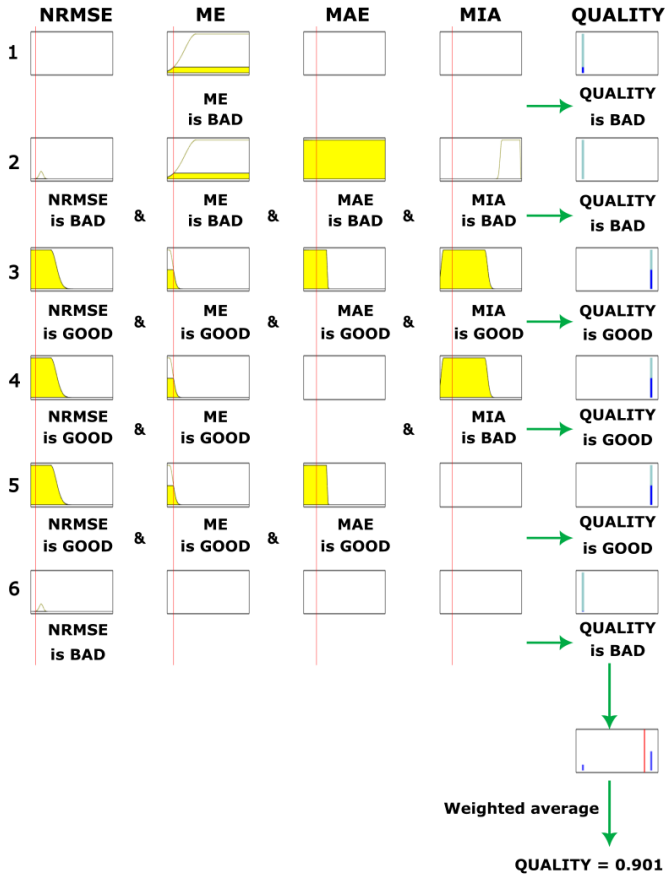


**Fig. 16.** TSK system with 4 inputs, 1 output and 6 IF-THEN rules. Each row represents one rule. The rules are presented as membership function. For rules combined with logical "AND" the minimum value from the certain rows is selected. After that the aggregation is performed based on the average mean value of the rules outputs

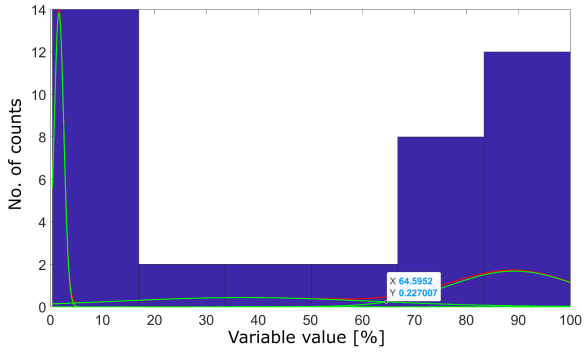


**Fig. 17.** VCMF Threshold defined by GMM decomposition with three components

## 4. RESULTS

For the models from the GOOD group, the VCMF box changed its position in comparison to the location of the box for the group of models not well-fitted to the data (Fig. 18). When comparing the locations of the boxes for different methods, it can be observed that only VCMF clearly differentiates the results between the GOOD and BAD groups. The median value is below 0.1 for

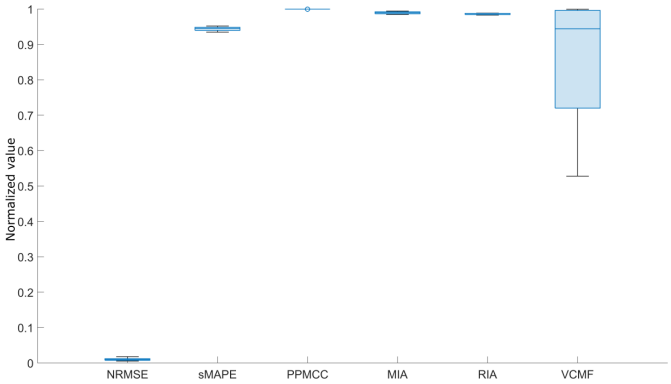


**Fig. 18.** Box plots for the models from the group GOOD for different goodness of fit evaluation methods (NRMSE, sMAPE, PPMCC, MIA, RIA) and VCMF

the GOOD group and over 0.9 for the BAD group. Figure 19 presents box plots for the group of poorly-fitted models (the group BAD). NRMSE and VCMF, MIA and RIA have similar median values. Only the VCMF box is high, meaning there is high variability within the results for different groups. The median value for the GOOD group is lower than 0.1, whilst for the BAD group, it is over 0.9. The obtained results proved that the proposed VCMF coefficient allows a better assessment of model quality compared to the mentioned measures of model quality assessment because it has a wide range of values, which allows differentiation into groups of models with a good fit to the actual measurement data and those with no good fit. Table 9 presents the values of the measures included in the VCMF coefficient. The first row indicates the results for correctly classifying the models into the BAD group by VCMF. The second and third rows contain models not accurately classified into the BAD group. It can be noticed that the impact of wrong classification has mainly the value of MAE. In the created TSH fuzzy-inference system, MAE has a very steep function, implicating a slight difference between the groups GOOD and BAD. In addition, the models that belong to the GOOD group include those for which the value of the MAE is less than 0.00300 and, at the same time, greater than 0.00270. In order to improve the quality of the VCMF coefficient, it would be necessary to modify the shape of the affinity function, which, however, will result in the group of GOOD incorrectly assigned models. The main functionality of the VCMF is the model goodness of fit evaluation. However, it can also automatically detect unforeseen plant behavior resulting in production failure. Using the digital twin technology, the model output can be constantly compared with the plant's actual output (Fig. 20). The proposed approach was also evaluated on a tank system shown in Fig. 21. The

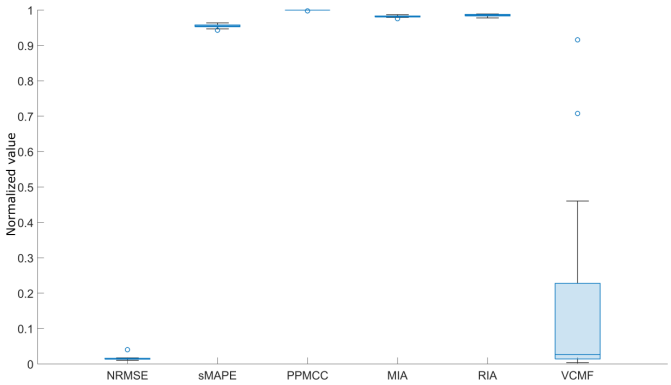


**Fig. 19.** Box plots for the models from the group BAD for different goodness of fit evaluation methods (NRMSE, sMAPE, PPMCC, MIA, RIA) and VCMF

**Table 9**
Values of measures included in the VCMF coefficient: first row – group *GOOD*, second and third rows – group *BAD*

| NRMSE | ME | MAE | MIA | VCMF |
|---|---|---|---|---|
| 0.01640 | 3.47620 | 0.00380 | 0.97930 | 0.000004 |
| 0.01060 | 5.26970 | 0.00240 | 0.98710 | 84.39650 |
| 0.01190 | 2.67250 | 0.00270 | 0,98610 | 80.60220 |

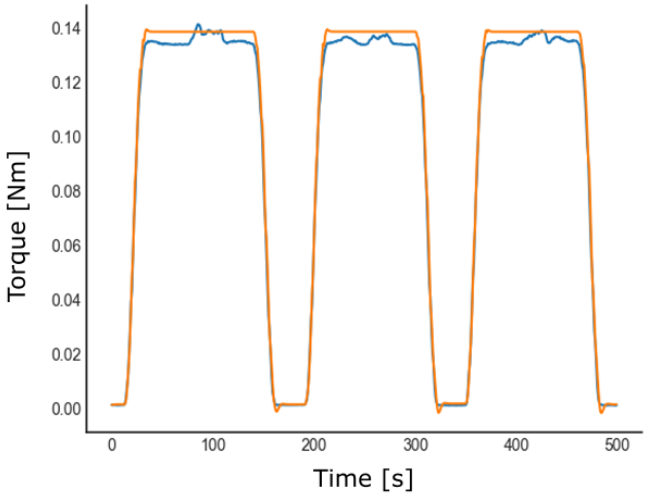


**Fig. 20.** Model goodness of fit – emergency situation detection. Blue – plant's actual output, orange – model output (fit to the data)

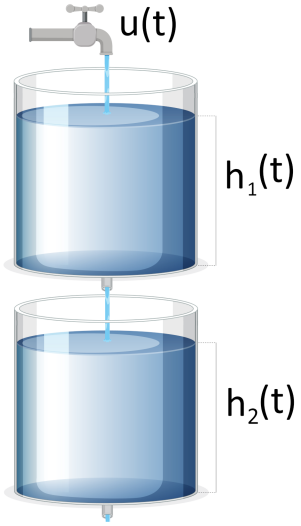


**Fig. 21.** Cascaded tank system

dataset consists of the normalized water level of the bottom tank and the normalized pump voltage. According to the box plot analysis, no outliers are present in the data (Fig. 22). The data do not follow a normal distribution, as indicated by the Shapiro–Wilk test with a p-value of $2.57 \times 10^{-36}$. An analysis based on Spearman's rank correlation reveals strong correlations between the variables (Fig. 23). Furthermore, the augmented Dickey–Fuller test confirms that the time series is stationary (test statistic: −2.87; *p*-value: 0.049). The seasonal component exhibits a constant amplitude, the trend component is approximately horizontal, and the remainder component represents random fluctuations in the time series (Fig. 24). The number of lags exceeding the confidence interval suggests a model order of 10 (Fig. 25). A pronounced peak is observed at lag 10, followed by a gradual decay at higher lags, indicating the presence of a moving-average (MA) component. Considering these observations, an ARMAX model with the pump voltage as an exogenous input was selected. The PACF diagram shows a significant peak at the first lag, followed by a gradual decay at higher lags (Fig. 26). The lag plots of the time series against its delayed values indicate linear relationships (Fig. 27), suggesting

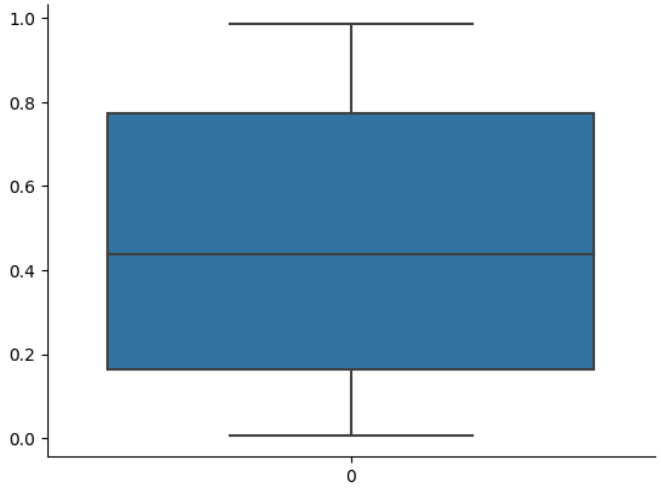


**Fig. 22.** Boxplot for two-tank system

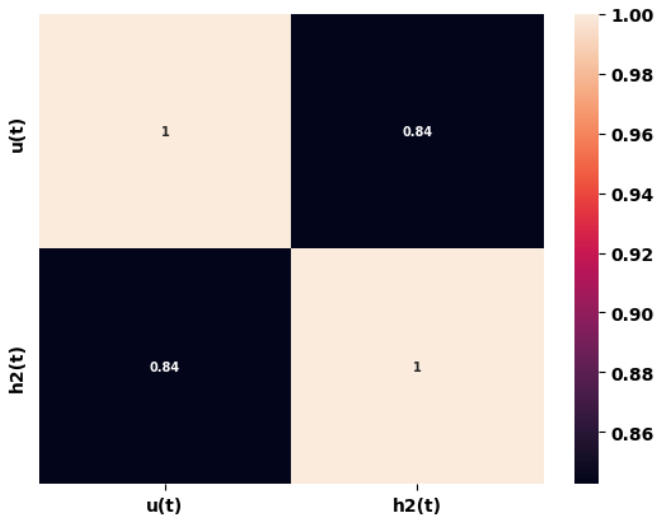


**Fig. 23.** Spearman's rank correlation between variables of the two-tank system

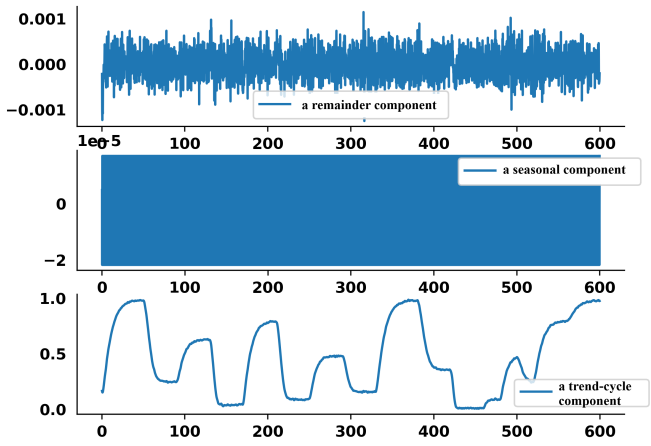


**Fig. 24.** The output of the time series decomposition

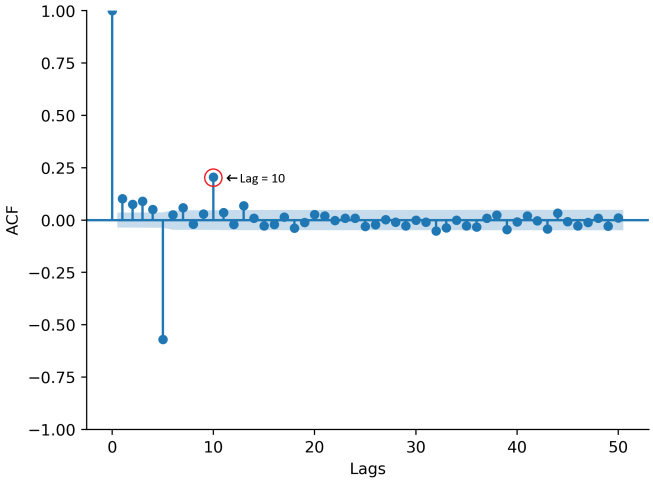


**Fig. 25.** Autocorrelation plot of the lower tank liquid level

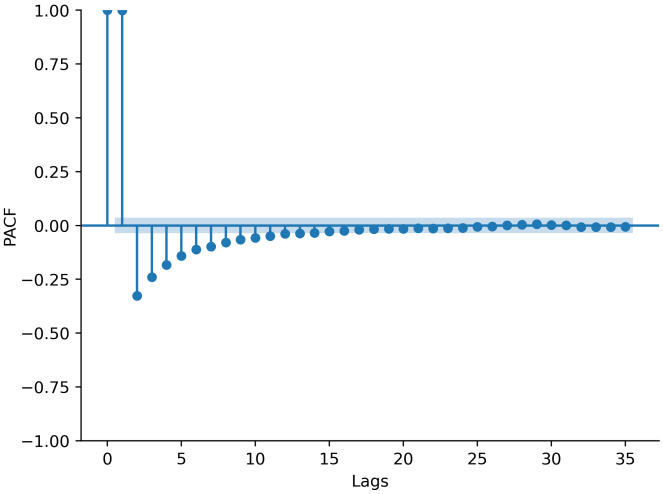


**Fig. 26.** Partial autocorrelation plot of the lower tank liquid level

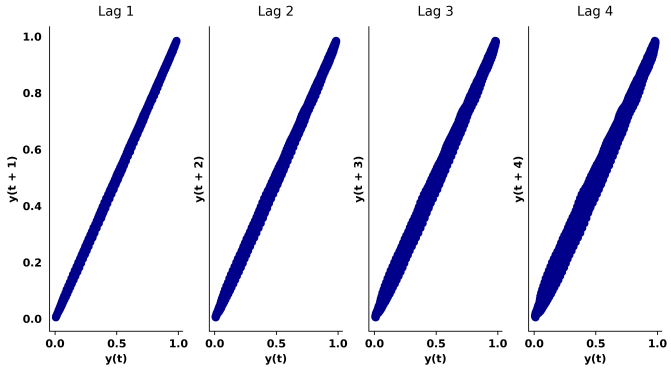


**Fig. 27.** Lag plots

that the data exhibit linear behavior. Therefore, the use of a nonlinear model is not required. Model parameters were selected using a neural network. Iteratively, 20 models with different parameters were created. Using the VCMF coefficient, the models were compared to each other – a model with coefficient values was selected. The VCMF coefficient was selected as the model representing the tanks' system. Table 10 shows the parameter values of the best model selected using the neural network. For the cascaded tank system, the VCMF coefficient was calculated to be 81.4%. Based on the cut-off value determined using the GMM algorithm, the obtained VCMF coefficient exceeds the threshold. This indicates that the system is described by a model that closely matches the actual data and is therefore suitable for virtual commissioning purposes (Fig. 28). For the belt conveyor models, the metric values and the VCMF coefficient were computed, and Wilcoxon's nonparametric test was applied because the data did not follow a normal distribution (Table 11). The null hypothesis stated that the data from the two groups (GOOD and BAD) originate from the same continuous distribution and have equal medians. The alternative hypothesis assumed that the data from the two groups do not come from the same continuous distribution and that their medians differ. Based on the obtained results, it can be observed that for RIA, ME, MSLE, and R2, the p-values exceed the significance level (0.05). This indicates that the medians of the two groups are not significantly different, leading to a failure to reject the null hypothesis. Consequently, these metrics are not capable of distinguishing between well-matched and poorly matched models. In contrast, VCMF, MIA, MAE, MedAE, EVS, PPMCC, sMAPE, MFE, RRSE, MSE, RMSE, and NRMSE yield p-values below the significance threshold, demonstrating their ability to differentiate between the GOOD and BAD model groups. This study introduced the Virtual Commissioning-Model-Fuzzy (VCMF) coefficient as a multi-criteria indicator for assessing model quality in virtual commissioning. By aggregating information from multiple goodness-of-fit metrics and incorporating expert knowledge through a Takagi–Sugeno–Kang fuzzy inference system, the proposed approach provides a more comprehensive and operationally meaningful evaluation than traditional single-metric methods. The experimental results obtained from a belt conveyor system and tests performed on a tank system demonstrate that VCMF effectively distinguishes well-matched models from poorly performing ones and offers improved interpretability for

**Table 10**
MAX model parameters

| $\theta_0$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|---|---|---|---|---|
| 3.3125 | 6.7677 | 10.4544 | 12.9922 | 13.4054 |
| $\theta_5$ | $\theta_6$ | $\theta_7$ | $\theta_8$ | $\theta_9$ |
| 11.5591 | 8.2346 | 4.6809 | 1.9573 | 0.4789 |

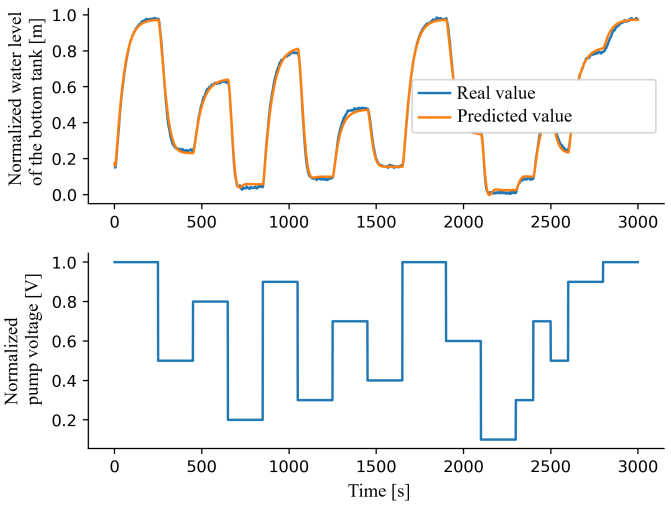


**Fig. 28.** Model fit to the measured data

**Table 11**
Wilcoxon's test results

| Name | $p$-value | $H_0$ reject? (1 – yes, 0 – no) |
|---|---|---|
| **VCMF** | $2.95975 \times 10^{-7}$ | 1 |
| MIA | $3.41558 \times 10^{-7}$ | 1 |
| MAE | $3.93881 \times 10^{-7}$ | 1 |
| MedAE | $7.94795 \times 10^{-7}$ | 1 |
| EVS | $1.20089 \times 10^{-6}$ | 1 |
| PPMCC | $1.37606 \times 10^{-6}$ | 1 |
| sMAPE | $2,06159 \times 10^{-6}$ | 1 |
| MFE | $3.06910 \times 10^{-6}$ | 1 |
| RRSE | $3.06910 \times 10^{-6}$ | 1 |
| MSE | $3.98735 \times 10^{-6}$ | 1 |
| RMSE | $3.98735 \times 10^{-6}$ | 1 |
| NRMSE | $4.54007 \times 10^{-6}$ | 1 |
| RIA | $0.28530 \times 10^{-6}$ | 0 |
| ME | $0.40935 \times 10^{-6}$ | 0 |
| MSLE | $6.67365 \times 10^{-6}$ | 0 |
| $R^2$ | $1.37606 \times 10^{-6}$ | 0 |

industrial practitioners. Despite these promising results, several constraints of the present investigation should be acknowledged. Some rare events, like emergency situations, or fault conditions should be more extensively examined in the future. Moreover, the VCMF coefficient was validated using offline data, whereas real-time or closed-loop VC environments may introduce additional challenges such as latency, noise, and operator interaction effects. To mitigate these limitations, future work should consider expanding the dataset to include multiple industrial processes with varying levels of complexity, enabling broader validation of the VCMF approach. In other industrial domains (e.g., robotics, process control) VCMF might perform similarly, but probably the small changes to the fuzzy rules or membership functions need to be introduced according to the expert's knowledge from different industrial scenarios or processes. Thus, future work should consider automation or employment of data-driven tuning methods and adaptive mechanisms. Furthermore, linking VCMF outputs directly to automated model selection, alarm filtering, or HMI visualization could significantly enhance the practical utility of virtual commissioning systems.

## ACKNOWLEDGEMENTS

The work was co-funded by POIR.01.01.01-00-0376/19 and SUT 02/050/BK26/0058.

## REFERENCES

[1] D. Chikko, M.J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than sMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, p. e623, 2021, doi: 10.7717/peerj-cs.623.

[2] J. Maedler, I. Viedt, and L. Urbas, "Applying quality assurance concepts from software development to simulation model assessment in smart equipment," in *Proc. 31st European Symposium on Computer Aided Process Engineering (ESCAPE31)*, vol. 1, 2021.

[3] J.M. Koehne, B.P. Mohanty, and J. Simuek, "Inverse Dual-Permability Modeling of Preferential Water Flow in a Soil Column and Implications for Field-Scale Solute Transport," *Vadose Zone J.*, vol. 5, no. 1, 2006, doi: 10.2136/vzj2005.0008.

[4] I. Viedt, J. Maedler, and L. Urbas, "Requirements for the quality assessment of virtual commissioning models for modular process plants," in *Proc. 14th International Symposium on Process Systems Engineering – PSE 2021+*, vol. 49, 2022, pp. 805–810, doi: 10.1016/B978-0-323-85159-6.50134-2.

[5] M. Segovia and J. Garcia-Alfaro, "Design, Modeling and Implementation of Digital Twins," *Sensors*, vol. 14, p. 5396, 2022, doi: 10.3390/s22145396.

[6] R. Ruzarovsky *et al.*, "Development and validation of digital twin behavioural model for virtual commissioning of cyber-physical system," *Appl. Sci.*, vol. 15, no. 5, p. 2859, 2025, doi: 10.3390/app15052859.

[7] O. Ismail *et al.*, "Systematic development of virtual commissioning architectures using spes methodology building blocks," *J. Intell. Manuf.*, 2025, doi: 10.1007/s10845-025-02759-2.

[8] Ł. Glodek, "Application of fuzzy systems for model quality assessment in virtual commissioning of industrial automation systems (Zastosowanie systemów rozmytych do oceny jakości modeli na potrzeby wirtualnego rozruchu systemów automatyki przemysłowej)," Ph.D. dissertation, Silesian University of Technology, 2023.

[9] S. Gramblička, R. Kohár, and M. Stopka, "Dynamic analysis of mechanical conveyor drive system," *Procedia Eng.*, vol. 192, pp. 259–264, 2017, doi: 10.1016/j.proeng.2017.06.045.

[10] M. Krzywinski and N. Altman, "Visualizing samples with box plots," *Nat. Meth.*, vol. 11, no. 2, pp. 119–120, 2014, doi: 10.1038/nmeth.2813.

[11] S.S. Shapiro and M. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965, doi: 10.1093/biomet/52.3-4.591.

[12] JMP, "Statistical discovery. how do i interpret the shapiro-wilk test for normality in JMP?" [Online]. Available: www.jmp.com/support/notes/35/406.html (Accessed 2023-02-24).

[13] J. Koronacki and J. Mielniczuk, *Statistics for Students of Technical and Natural Sciences (Statystyka dla studentów kierunków technicznych i przyrodniczych)*. Warsaw: Wydawnictwo Naukowo-Techniczne, 2006.

[14] C. Spearman, "The proof and measurement of association between two things," *Am. J. Psychol.*, vol. 15, pp. 72–101, 1904, doi: 10.1093/ije/dyq191.

[15] C. Spearman, "Footrule for measuring correlation," *Brit. J. Psychol.*, vol. 2, pp. 89–108, 1906, doi: 10.1111/j.2044-8295.1906.tb00174.x.

[16] P. Schober, C. Boer, and A. Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, 2018, doi: 10.1213/ANE.0000000000002864.

[17] G. Kirchgässner and J. Wolters, *Introduction to Modern Time Series Analysis*. Berlin, Heidelberg: Springer-Verlag, 2007.

[18] GUS, "Time series (szeregi czasowe)." [Online]. Available: https://stat.gov.pl/metainformacje/szeregi-czasowe-4712/ (Accessed 2023-03-03).

[19] R.J. Hyndman and G. Athanasopoulos, "Principles and practice in OTexts, 3rd ed. OTexts 2024." [Online]. Available: https://otexts.com/fpp3/

[20] X. Wang, Y. Hua, E. Kodirov, D. Clifton, and N. Robertson, "Imae for noise-robust learning: Mean absolute error does not treat examples equally and gradient magnitude's variance matters," *arXiv:1903.12141*, 2019, doi: 10.48550/arXiv.1903.12141.

[21] C. Tianfeng and R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.

[22] S.N. Sivanandam, S. Sumathi, and S.N. Deepa, *Introduction to Fuzzy Logic using Matlab*. Berlin, Heidelberg: Wiley Online Library, 2010.

[23] Ł. Glodek, S. Bysko, and W. Nocoń, "Model goodness of fit for virtual commissioning purposes based on fuzzy-inference system," in *ISEEIE 2021: 2021 International Symposium on Electrical, Electronics and Information Engineering*, 2021.

[24] J. Łęski, *Neuro-Fuzzy Systems (Systemy neuronowo-rozmyte)*. Warsaw: Wydawnictwo Naukowo-Techniczne, 2008.

[25] D. Dubois and H. Prade, "The three semantics of fuzzy sets," *Fuzzy Sets Syst.*, vol. 90, no. 2, pp. 141–150, 1997, doi: 10.1016/S0165-0114(97)00080-8.

[26] The MathWorks Inc., "Statistics and machine learning toolbox," 2022. [Online]. Available: www.mathworks.com/help/stats/index.html (Accessed 2023-05-10).

[27] A. Nikolaev and S.H. Jacobson, "Simmulated annealing," in *Handbook of Metaheuristics*, F. Glover and G.A. Kochenberger., Eds. Springer, 2010, doi: 10.1007/978-1-4419-1665-5_1.

[28] J. Cohen, *Statistical power analysis*. New York: Erlbaum, 1988.

[29] J. Frost, "Statistics by Jim." Making statistics intuitive. [Online]. Available: https://statisticsbyjim.com/glossary/pooled-standard-deviation

[30] A. Polański, M. Marczyk, M. Pietrowska, P. Widłak, and J. Polańska, "Signal Partitioning Algorithm for Highly Efficient Gaussian Mixture Modeling in Mass Spectrometry," *PLOS ONE*, vol. 10, 2015, doi: 10.1371/journal.pone.0134256.

[31] D.F. Findley, "Counterexamples to parsimony and BIC," *Ann. Inst. Stat. Math.*, vol. 43, pp. 505–514, 1991, doi: 10.1007/BF00053369.