International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I8P129

Volume 72 Issue 8, 312-324, August 2024

Original Article

New Hybrid Weighted Predictive Data Analysis Approach: Application to Temporal Power Transformer Maintenance Data

Moïse Manyol[1*], Samuel Eke[2], Georges Olong[3], Ndoumbè Matéké Max[4], Alain Biboum[5], Ruben Mouangue[6]

[1,2,3,4,6]*Electronics, Electrotechnics, Automation and Telecommunication Research Laboratory (LEEAT), Higher National Polytechnic School, University of Douala, Douala/Cameroon,*
[5]*Mechanical and Industrial Engineering Department, National Advanced School of Engineering of Yaounde, University of Yaounde I, Yaounde/Cameroon.*

[1]*Corresponding Author : moisemany@yahoo.fr*

Received: 29 April 2024 Revised: 10 July 2024 Accepted: 31 July 2024 Published: 28 August 2024

Abstract *- This paper proposes a new methodological approach for extracting typical data profiles from time series of power transformer maintenance databases with the aim of preventing future faults. The proposed approach operates in two stages. First, the model formalizes the date variable as a time series and then analyzes the data to identify interdependencies between them. Rio Tinto Alcan's dissolved gas data from transformer T0001 show that H2, CH4, C2H4, C2H6, CO, CO2, and N2 are dependent, while C2H2 and O2 are independent. In the second step, to capture short- and long-term trends, seasonalities, and dependencies in the data on the one hand and to extract non-linear trends and seasonalities for variable forecasting on the other, an ARIMA (Autoregressive Integrated Moving Average) + GES (Generalized Exponential Smoothing) model in combination is applied to the data series. The hybrid ARIMA (2, 1, 1) + GES (0.1, 0.1, 1) model, with a weighting of 0.5, produced errors of 0.27 and 4.5%, respectively, in terms of Mean Absolute Scaled Error (MASE) and Mean Symmetric Absolute Percentage Error (SMAPE). The ARIMA model taken individually gave MASE equals 0.55 and SMAPE equals 8.9%. Similarly, the proposed model is better than the naive model because the MASE is less than 1 (0.27%). The data series was subjected to other forecasting models, and it was found that the model proposed in this article is more accurate given the error results obtained since the smaller the error, the more accurate the forecasting model.*

Keywords *- Smoothing, Autoregression, Seasonality, Trend, Forecasting.*

1. Introduction

In the electrical energy production chain, the transformer is one of the most important components, and its failure would have a major impact on the country's economy and its consumers. On September 20, 2022, the Logbaba transformer station in Douala, Cameroon, caught fire, depriving the industrial zone of electricity. In December 2013, a similar disaster occurred in New Jersey, USA, where around 12,000 people lost their source of electrical power due to a fault in the transformer [1]. Likewise incident took place in Stamford, in the USA, in which a transformer caught fire, depriving over 1000 people of light for days. In view of the above, it becomes imperative to adopt new attitudes to prevent future failures of power transformers. The analysis of data from power transformers is today the subject of much research. Thus, to detect and predict transformer faults, the work [3] designed a hybrid system to control maintenance. the system uses a genetic algorithm and a neural network. The genetic algorithm was used to cluster gas entry concentrations, and the neural network was used to predict faults present in the transformer by generating decision rules. In 2023, lightning failure data was used to Predict transformer failures in the works [4]. A Single-class hybrid Vector Deep Data Description (SVDD) that uses the Synthetic Minority Oversampling (SMOTE) Technique to manage data misbalance between the minority and majority class tags is used. The Maximum Relevance Minimum Redundancy (MRMR) is used as a feature selection technique to improve model accuracy. This model was compared to five benchmark models. to prevent and classify power transformer failures, the work [5] used different failure classification techniques based on dissolved gas analysis data, mainly logistic regression, multiclass jungle, multiclass decision tree, and artificial neural network. The application can diagnose power electrical transformer failures based on the parts per million of the various gases generated in the oil. Work [6] has developed a criterion for the dimensioning transformers in railway systems based on the moving average of apparent power.

The results obtained with this method were compared with reference designs obtained using standard thermal criteria. This comparison reveals that moving-average methods facilitate the evaluation of designs with an uncertainty of accuracy. Peimankar et al. (2018) developed three approaches to the multi-objective ensemble for predicting dissolved gas contents in electrical power transformers: MOPSO-based ensemble time-series forecasting, NSGA-II-based ensemble time-series forecasting, and SPEA-II-based ensemble time-series forecasting.

In addition, these methods were compared to four different techniques, namely Autoregressive Integrated Moving Average (ARIMA), Simple Exponential Smoothing (SES), Persistence Model (PER) and Weighted Ensemble Method (WENS), which assigns a normalized weight to the ensemble data. It is important to note that the model proposed in this study also has certain limitations. For example, the main application of machine learning techniques to time series forecasting tasks is to select the most appropriate lags in the time series as inputs for the forecasting models [7].

In 2011, the Artificial Neural Network (ANN) classifier and exponentially weighted moving averages were used in an asset management framework for a power company's power transformers. After training the ANN, control limits were established using the EWMA (Exponentially Weighted Moving Average graph) method. This graph was generated from the error data of the validation set, giving a weight of 0.9 and multiples of 3 of the standard deviation.

However, given the complexity of the practical implementation of the proposed model, a simple approach to maintenance scheduling using asset prioritization diagrams is also proposed as an alternative to support decision-making [8]. In 2016, as part of the analysis of water consumption based on meter data, a regression based on a series of kilns and moving averages was used to extract typical consumption profiles [9]. In 2022, to detect anomalies in vehicle charging stations, data-driven thermal modeling was based on the combination of an absolute error measure and an exponential moving average filter.

This enabled anomalies to be detected more reliably than more advanced measures such as the Mahalanobis distance; other types of filters, such as the simple moving average, work very similarly to the exponential moving average, provided their parameters are set appropriately [10]. This method has proved accurate and reliable on simulation data but requires further work to verify its capabilities in a more realistic scenario. To monitor and prevent the lifetime of power transformers, work [11] has adopted the weighted moving average. Indeed, the Weighted Moving Average, whose trend is taken over the year, is not affected by short-term fluctuations, unlike the exponential Weighted Moving Average, and it is easier to use the MMP model than regression models in the programming phase. In 2019 to assess failure rates from transformer data, Moving Averages are used to present trends in failure rates for different transformer types and sizes. This work provides a context for the application of dynamic failure data to Reliability Centered Maintenance (RCM) principles and beyond. It shows that equipment age and maintenance affect component failure despite the non-exhaustive database [12]. The above review shows that time series analysis is widely used for fault prediction. Thus, power transformer failure prediction is the subject of much attention in the scientific community today.

This paper proposes a new hybrid weighted regression approach based on the exponential smoothing technique and the ARIMA regression model. This approach begins with formatting maintenance time data into chronological data using one-hot encoding. The ARIMA model combines autoregressive (AR- Autoregressive), integrated (I- Integrated), and moving average (MA- Moving Average) processes [13, 14]. Exponential smoothing here, which is of three orders: Simple, Double, and Generalized [15], takes account of trends and seasonal patterns in the data by using a smoothing constant that weights the most recent observations against the older ones. As for the ARIMA regression model, it also takes into account seasonal differences, trends, and the effects of recent data history to provide more accurate and comprehensive forecasts.

By combining the ARIMA model of a certain rank with the generalized exponential smoothing model and with a weighting of order less than unity, it is possible to minimize information loss while providing accurate trends and forecasts for the data. Dissolved gas data from Rio Tinto Alcan's T0001 transformer were used as an application. Following a review of exponential smoothing and ARIMA fitting techniques in Section 2, the proposed regression approach is presented in Section 3. Finally, the results of our method applied to a dissolved gas database are compared with other methods presented in Section 4, followed by a discussion.

2. Materials and Methods

The methods and materials presented here are those used in the proposed approach. These are mainly moving-average-based adjustment techniques. Here, we present some of the techniques commonly used in adjustment and forecasting models.

2.1. Moving Averages

This tool is an indicator reflecting the average valuation of a stock over a given period. Moving averages are very simple in principle, do not a priori require the use of sophisticated concepts or models, and are particularly flexible in application. By definition, a moving average is equal to the weighted sum of values of X corresponding to dates surrounding t. A moving average of order m_1+m_2+1 is written as:

$$X_t^* = \sum_{i=-m_1}^{m_2} \theta_i X_{t+i} = \theta_{-m_1} X_{t-m_1} + \theta_{-m_1+1} X_{t-m_1+1} + \cdots + \theta_{m_2} X_{t+m_2} \quad (1)$$

Where

$$m_1 \geq 0,\ m_2 \geq 0,\ \theta_i \in \Re$$

And θ the vector of dimension ($m_1 + m_2 + 1, 1$) whose coordinates are the coefficients of the moving average:

$$\theta = \begin{bmatrix} \theta_{-m_1} \\ \theta_{-m_1+1} \\ \vdots \\ \vdots \\ \theta_{m_2-1} \\ \theta_{m_2} \end{bmatrix}$$

The different types of moving averages and their properties have been extensively detailed in works [16, 17, 18]. Here, we present some of them and their composites, which we consider relevant.

2.2. The ARIMA Model

The ARIMA class of models was introduced to reconstruct the behavior of processes subject to random shocks over time. They combine three types of temporal processes: autoregressive (AR-AutoRegressive), integrated (I-Integrated), and moving average (MA-Moving Average).

The contribution of each is specified by the notation ARIMA(p,d,q), where *p* is the order of the AR(p) autoregressive process, d is the degree of integration of an I(d) process, and q is the order of the MA(q) moving average [14]. As a result, several research studies have used these models to identify a suitable autoregressive integrated moving average model and a suitable state-space model for a time series such as [19], with results showing that these models consistently provide more accurate forecasts than other approaches and that the improvements in accuracy are significant.

2.2.1. Autoregressive Processes (AR)

For an autoregressive process, each value in the series is a linear combination of the previous values in the series. An autoregressive process in which the value of the series at time *t*, X_t, depends on the p previous values within a random perturbation ε is said to be of order p and denoted AR(p). The process is thus written:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{n} \beta_i X_{i-1} \quad (2)$$

β is the Autoregression coefficient and expresses the strength of the linear link between two successive values. An autoregressive process can be said to have a "memory" in the sense that each value is correlated with the set of values preceding it.

2.2.2. Integrated Processes (I)

The mean of the ARIMA model series is constant over time, as is the variance (stationary). To eliminate any tendency, it is important to differentiate, i.e. to replace the original series with the series of adjacent differences. A time series that needs to be differentiated to achieve stationarity is considered to be an integrated version of a stationary series (hence the term Integrated). An integrated *I(1)* process of order 1 is written as : ARIMA(*0,1,0*) :

$$X_t = \mu + X_{t-1} + \varepsilon_t \quad (3)$$

The second-order models work on difference differences and no longer on raw differences. The second difference of *X* at time t is defined by:

$$(X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \rightarrow X_t - 2X_{t-1} + X_{t-2} \quad (4)$$

ARIMA (*0,2,0*) will obey the following predictive equation:

$$X_t - 2X_{t-1} + X_{t-2} = \mu + \varepsilon_t \quad (5)$$

Where the random disturbance εt is white noise, and μ is the model constant, and represents the mean difference in *X*. If μ is *0*, the series is stationary.

2.2.3. Moving Average (MA)

Moving average models suggest that the series fluctuates around a mean value. We then consider that the best estimate is represented by the weighted average of a certain number of previous values (which is the principle of moving average procedures used for data smoothing).In effect, this amounts to considering that the estimate is equal to the true average, to which we add a weighted sum of the errors that have marred the previous values:

$$X_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} \ldots . + \varepsilon_t \quad (6)$$

Typically, each observation is characterized by a random error component (ε) and a linear combination of past random errors. θ_1, θ_2 and θ_3 are the moving average coefficients of the model [20, 21].

2.3. Exponential Smoothing

Introduced by Holt in 1958 and Winters in 1960 and popularized by Brown's book (1963), smoothing methods constitute the set of empirical forecasting techniques which give more or less importance to the past values of a time series. As in the moving-average method, each piece of data is successively smoothed, starting from the initial value. Exponential smoothing gives past observations a weight that decreases exponentially with their age. There are many different smoothing techniques, some of which are described below.

2.3.1. Simple Exponential Smoothing (SES)

This method is used to forecast a time series when there is no trend or seasonal pattern, but the mean of the time series

X_t evolves slowly over time. The weighting is exponential so that more recent observations are given greater weight than older ones. This method is mainly applied to short-term forecasts, i.e. generally for periods not exceeding one month [22, 20].

$$Y_t = \alpha X_t + (1-\alpha)Y_{t-1} \quad (7)$$

The parameter α is a smoothing factor between *0* and *1*. In other words, Y_t can be seen as a weighted average between the current value X_t and the previous smoothed value Y_{t-1}.

2.3.2. Double Exponential Smoothing (DES)

Simple exponential smoothing does not give good results when the raw data show a trend or trends. The smoothed values are systematically underestimated or overestimated, depending on the direction of the trend. The purpose of double exponential smoothing methods is to smooth the level of the data (i.e. eliminate random variations) and to smooth the trend, i.e. eliminate the effect of the trend on the smoothed values. There are two methods of double exponential smoothing: Holt's method extended by Winters and Brown's method [22].

$$Y_t = \alpha X_t + (1-\alpha)(Y_{t-1} + T_{t-1}) \quad (8)$$

$$T_t = \beta(Y_t - Y_{t-1}) + (1-\beta)T_{t-1} \quad (9)$$

The data begins at time $t = 0$; again, X_t is the raw data series. At least X_0 and X_1 are available. The term Y_t is the series of smoothed values, and Tt is the trend estimated. The choice of the initial value X_0 is a matter of practice; we can take as a starting point an average of a number of previous past values. Starting from period *t*, a forecast for period $t + m$ is given by:

$$P_{red_{t+m}} = Y_t + mT_t \quad (10)$$

The level (Y_t) is predicted by the value of the levelled data at the end of each period. The trend (T_t) is predicted by the average levelized increase at the end of the period (Dhamodharavadhani and Rathipriya, 2019).

2.3.3. Generalized Exponential Smoothing (GES)

Generalized exponential smoothing is an even more flexible method that takes into account both trends and seasons in the time series. It uses several parameters that can be adjusted to suit the specific characteristics of the time series, such as trend, seasonality, level changes and error effects. The generalized exponential smoothing equation and the trend at time t can be estimated as follows:

$$Y_{t+1} = \alpha Y_t^{\gamma} + (1-\alpha)(Y_t + T_t) \quad (11)$$

$$T_t = \beta(Y_t - Y_{t-1}) + (1-\beta)T_{t-1} \quad (12)$$

Where:

α *and* β are the smoothing constants $\epsilon\ ]0,1[$

γ is the shape parameter ($\gamma > 0$)

2.4. Model Evaluation

To optimize the forecast, whether it is a moving average, exponential smoothing, or another form of forecasting, it is imperative to calculate and evaluate MAE, MASE, MSE, RMSE, and SMAPE [23] for a time series with n observations. Let us denote by y_i the historical observation at the given time and by y_p the forecast.

The Symmetrical Mean Absolute Percentage Error (SMAPE) is an alternative to the Mean Absolute Percentage Error (MAPE) when there are zero or near-zero values in the actual observations. SMAPE is self-limiting, reducing the influence of zero or near-zero observations. SMAPE is the forecast minus the actuals divided by the sum of the forecast and actuals as expressed in this formula:

$$SMAPE = \frac{2}{n}\sum_{i=1}^{n}\frac{|y_i - y_p|}{|y_i| + |y_p|} \quad (13)$$

Mean Absolute Error (MAE) Corresponds to the mean absolute difference between the model-adjusted values (forecast in the model one step ahead) and the observed historical data. It measures the mean of the residuals in the data set.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_p\right| \quad (14)$$

Mean Absolute Scaled Error (MASE) is the error measure used for model accuracy. It is the MASE divided by the MASE of the naive model.

The naive model predicts the value at a time point t as the previous historical value. Scaling according to this error means you can assess the quality of the model compared to the naive model. If the (MASE) is greater than 1, the model is worse than the naive model. The lower the MASE, the more qualitative the model is compared to the naive model.

$$MASE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - y_p|}{MAE^*} \quad (15)$$

Where MAE^* can be defined as naive forecasts for nonseasonal time series [24].

Root Mean Squared Error (RMSE) It is on the same scale as the observed data values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_p)^2} \quad (16)$$

3. A Proposed Approach to Data Analysis

Figure 1 shows a technique for analyzing power transformer maintenance data that combines the ARIMA model and generalized exponential smoothing techniques with prior pre-processing. The interest of such a technique lies in its combined advantages in terms of speed, robustness, and accuracy.

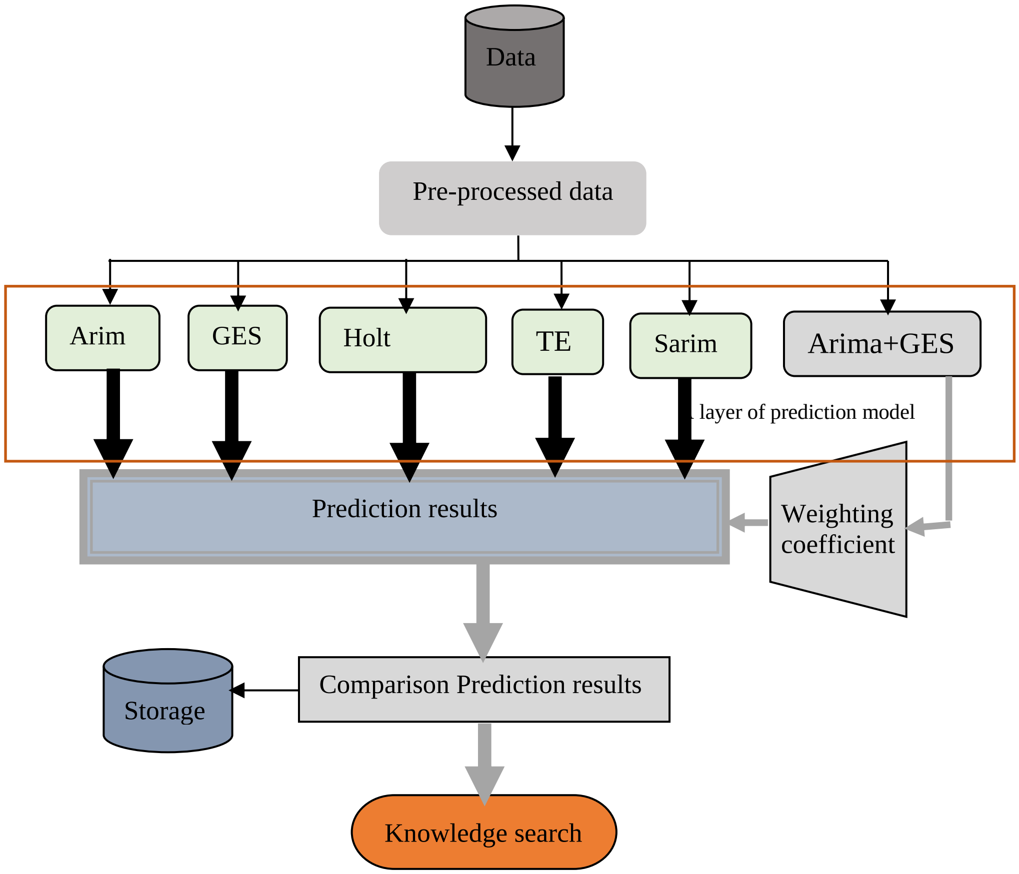


Fig. 1 Technique for analyzing power transformer

3.1. Algorithm Presentation

The first step is to compile a table of maintenance data for power transformers. For this, we have a database from Rio Tinto Alcan of Canada, in which we will exploit the GD (dissolved gas) data of the T0001 equipment. Pre-processing is applied to this database. The ARIMA, VARIMA, Holt-Winter, GES, and MA models form the predictive analysis layer. To make the series data stationary, the ARIMA model incorporates differentiation techniques, and subsequently, an adjustment can be applied to model the data series. Generalized exponential smoothing is a forecasting method that uses exponential weights to give more weight to more recent observations. GES can model seasonal patterns, no linear trends and no-Gaussian errors. GES can, therefore, be used as an alternative or complement to ARIMA to model complex series and thus have a robust predictor [2].

3.2. Differentiation Techniques

Several differentiation techniques are used to make a series stationary. Stationarity is important because it simplifies the statistical properties of the series, making forecasting and analysis easier. The main differentiation techniques used in the ARIMA model to make a series stationary are:

3.2.1. First-Order Differentiation

This involves taking the difference between each observation and its previous observation. If the time series has a trend or seasonality, first-order differentiation can help eliminate them.

3.2.2. Seasonal Differentiation

This involves taking the difference between each observation and the observation at the same time in the previous season. If the time series has seasonality, the seasonal difference can help eliminate it.

3.2.3. Additional Differentiation

If the time series is not yet stationary after initial and/or seasonal differentiation, further differentiation may be necessary. This involves taking the difference between each observation and its previous observation after the first and/or seasonal differentiation. Let ∇ be the differentiation operator, i.e. the operator defined by:

$$\nabla X_t = X_t - X_{t-1} \quad (17)$$

For all $t \geq 2$.

The recurrence formula defines the differentiation operator of order k.

$$\nabla^{(k)} X_t = \nabla(\nabla^{(k-1)} X_t \quad (18)$$

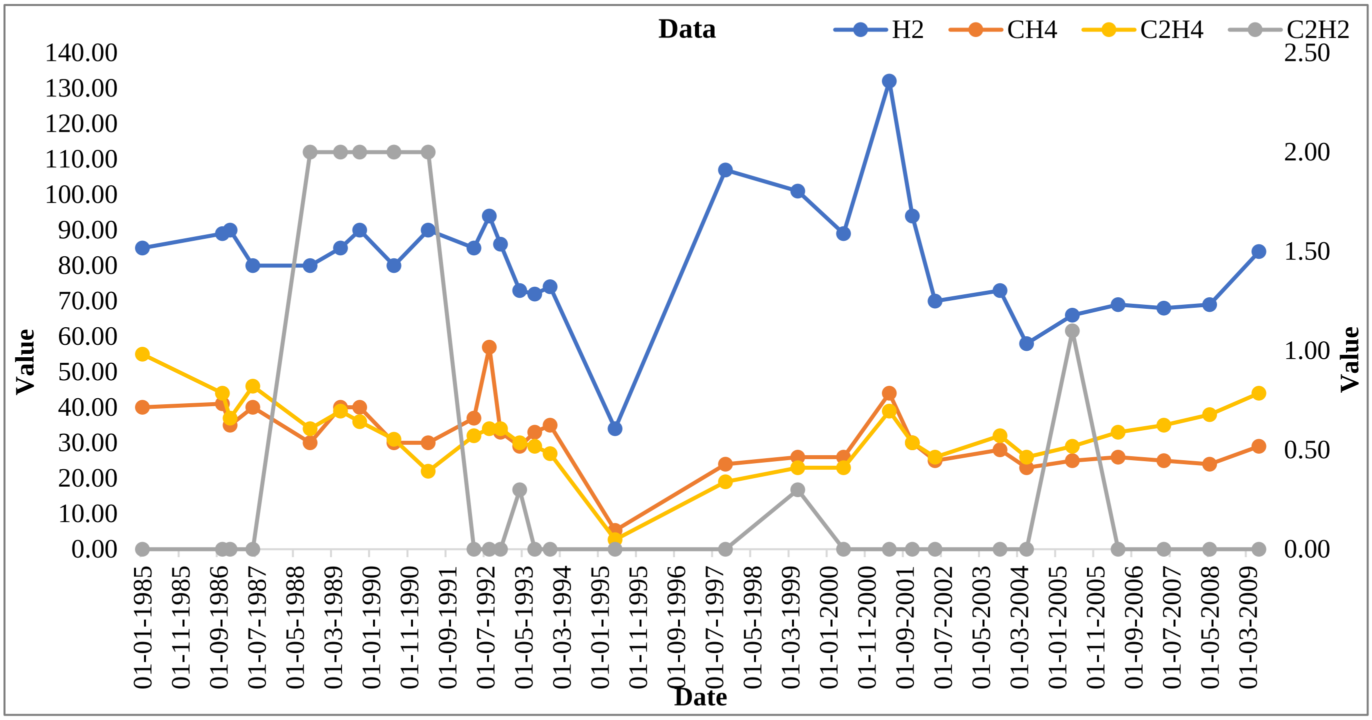


Fig. 2 RIO Tinto Alcan of Canada database of dissolved gases from transformer T0001

Table 1. Data processing statistics

Variable	Observations	Obs with missing value	Obs without missing value	Minimum	Maximum	Mean	Standard deviation
H2	29	0	29	34.000	132.000	81.621	17.174
CH4	29	0	29	5.300	57.000	31.390	9.194
C2H2	29	0	29	0.000	2.000	0.403	0.772
C2H4	29	0	29	2.700	55.000	32.059	9.679
C2H6	29	0	29	2.700	52.000	33.472	13.959
CO	29	0	29	242.000	1929.000	1207.690	244.663
CO2	29	0	29	1181.000	15092.000	9413.690	2733.602
O2	29	0	29	403.000	19996.000	5198.621	4502.169
N2	29	0	29	58747.000	105155.000	70573.586	10666.770

4. Results and Discussion

In this section, we present the results obtained after data analysis based on the generalized exponential smoothing model with optimal coefficient, then on the multivariate ARIMA model hybridized to the GES, and finally, a comparison of the results obtained from the different models will be discussed. Figure 2 shows our database with the variables hydrogen (H_2), methane (CH_4), ethylene (C_2H_4), and acetylene (C_2H_2). Looking at these data, we can see that the 'Date' variable shows a somewhat non-linear succession of dates. For the rest of the analysis, it is therefore important to make this variable chronological.

4.1. One-Hot Encoding

It is possible to convert a non-chronological variable into a chronological one using certain techniques. One-hot encoding represents each category of a categorical variable as separate binary variables. If a categorical variable has N unique categories, one-hot encoding will create N new binary variables, where each variable corresponds to a specific category. To represent a given category, a binary variable will take the value 1 if the observation belongs to that category and 0 otherwise. Consequently, only one variable will be active (1) for each observation, while all other variables will be inactive (0). The main structure of the algorithm is shown below.

» # 1. Define the list of possible categories
» # 2. Create a matrix of zeros
» # 3. For each observation i:
 # a. Find the index k corresponding to the category of observation i
 # b. Set a value of 1 in row i and column k of the matrix
» # 4. Return the encoded matrix

4.2. Data Analysis

A correlation is a statistical tool used to analyze linear relationships between different variables. It provides a correlation matrix that measures the strength and direction of the relationship between pairs of variables. Table 1 Above presents a descriptive analysis of the statistical data, highlighting their quality. Figure 3 shows that the variables O_2 (oxygen) and C_2H_2 (acetylene) are independent, with correlations between 0.01 and 0.3, values very close to unity. On the other hand, CO_2 (carbon dioxide) and C_2H_6 (ethane) are highly dependent. Clearly, CO_2 and C_2H_6 are largely influenced by butane (CH_4), ethylene (C_2H_4) and carbon monoxide (CO).

4.3. Results of the Developed Hybrid Model

The hybridization principle is based on the use of an additive ARIMA model and generalized exponential smoothing. The model is then evaluated.

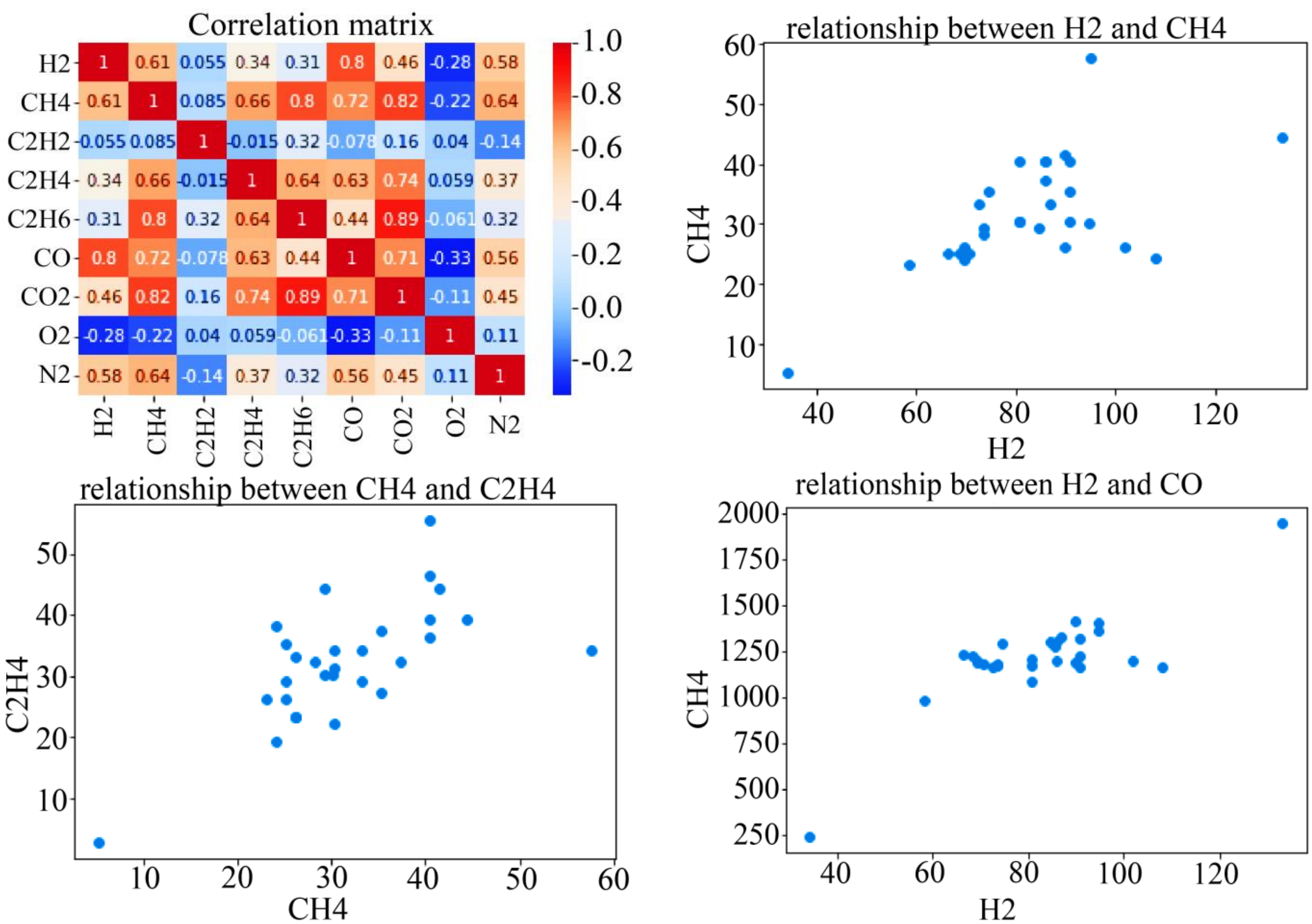
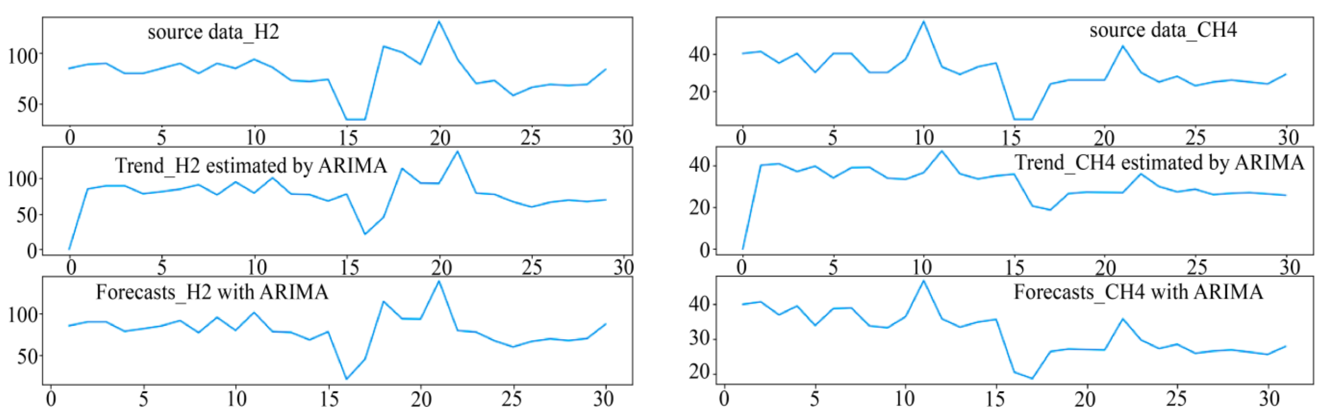


Fig. 3 Pearson correlation matrix



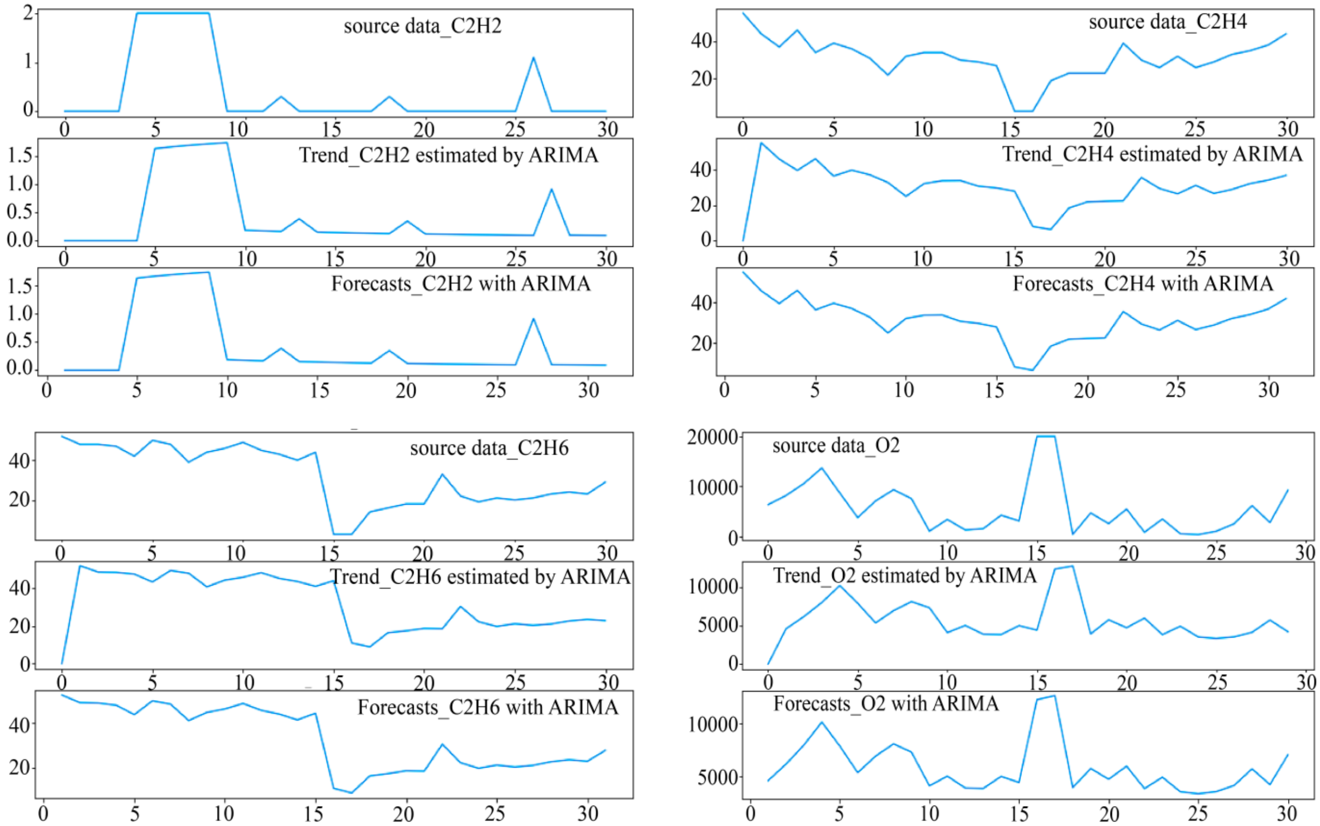


Fig. 4 Trend and forecast curves for variables

4.3.1. Predictive Analysis Using the ARIMA Model

After several observations, we decided on the ARIMA (2, 1, 1) model. This uses two AR terms, a differentiation of order 1 and an MA term. This means it captures the linear relationship between one observation and the two preceding observations, differentiates the series once to make it stationary, and also incorporates a linear dependence on the errors of the lagged observations. The prediction results for the different variables are shown in Figure 4. The resulting observation shows that, despite the slight differences in amplitude, the prediction curves are almost identical to the original data curves, implying that the model chosen is accurate.

4.3.2. Predictive Analysis Using Generalized Exponential Smoothing

This technique is an extension of exponential smoothing, enabling us to take into account more complex models and data with different properties. Additive GES (Gaussian Exponential Smoothing) uses exponential weights to calculate future predictions based on past observations. These weights are determined by specific parameters that are adapted to each data series. After fitting alpha(α), beta(β), and gamma(γ), the generalized approach was able to model and capture various properties of the data, such as seasonality, trend, prediction curve, and level for all the variables in the table. What was retained as fitting parameter values after several trials are alpha ($\alpha = 0.1$), beta($\beta = 0.1$), and gamma($\gamma=1$). The results of the prediction of hydrogen (H_2) and methane (CH_4), gases produced at low temperatures during the decomposition of mineral oil, are shown in Figure 5.

To avoid underestimating or overestimating the seasonal effect, gamma has been set to 1. This setting gives balanced importance to trend and seasonality in the forecasts. By adjusting the gamma parameter to other values, we can give more or less relative importance to the seasonal component compared with the trend. Figure 5 shows that the seasonality curves follow the shape of the forecast curves and historical dissolved gas data from power transformer maintenance. This result indicates that seasonal variations have been correctly taken into account in the forecast and that the model used is reliable. It also suggests that seasonal trends have been well analyzed and that the forecast is in line with the existing model. Overall, this indicates increased confidence in the reliability and accuracy of the forecasts. Combining the ARIMA (AutoRegressive Integrated Moving Average) model with Generalized Exponential Smoothing, the results obtained by taking one of the variables in the data series, such as hydrogen (H2), are shown in Figure 6. This prediction curve (GesAri2) tends to follow closely that of the original data, meaning that the prediction model is accurate to within a small margin of error.

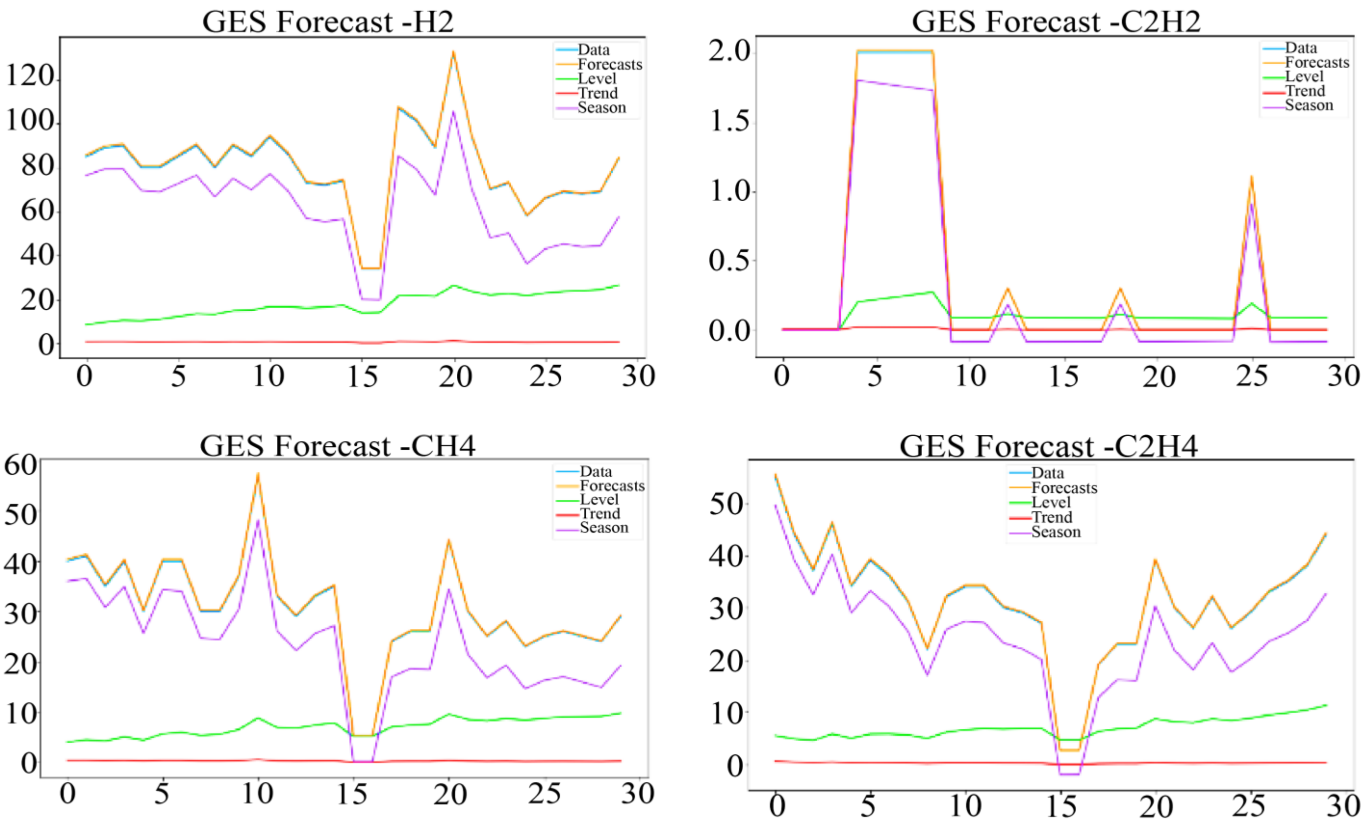


Fig. 5 GES model forecast results

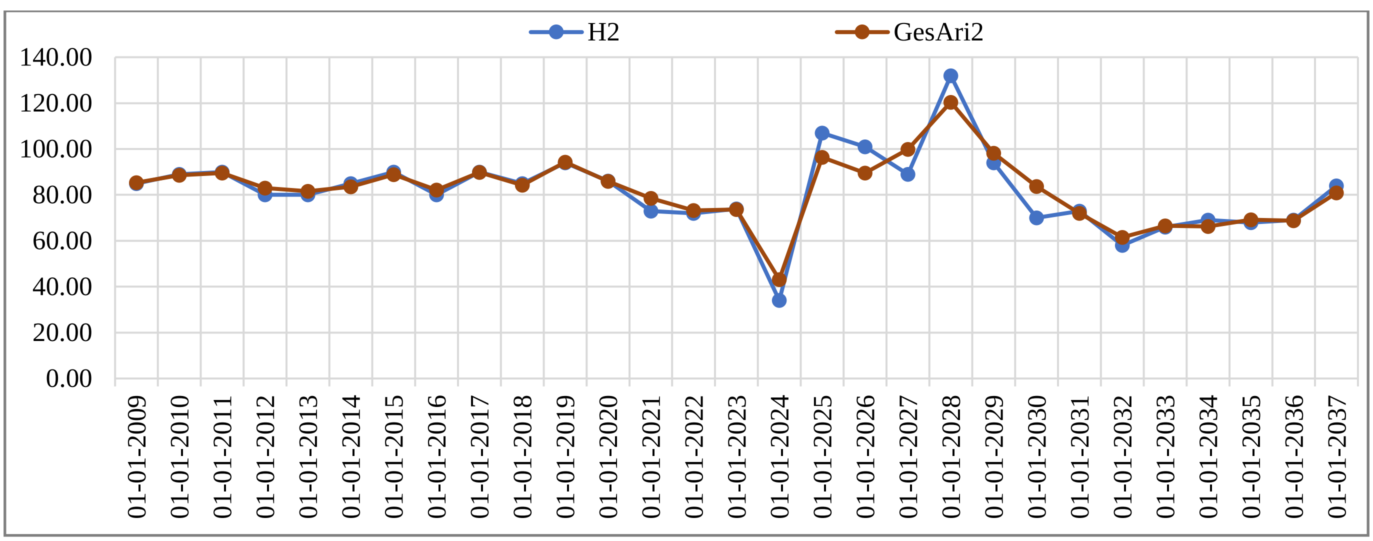
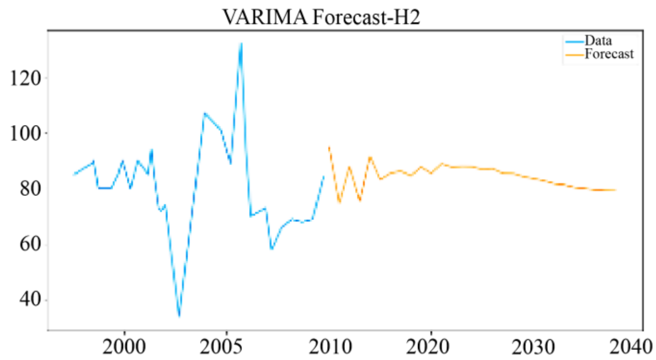


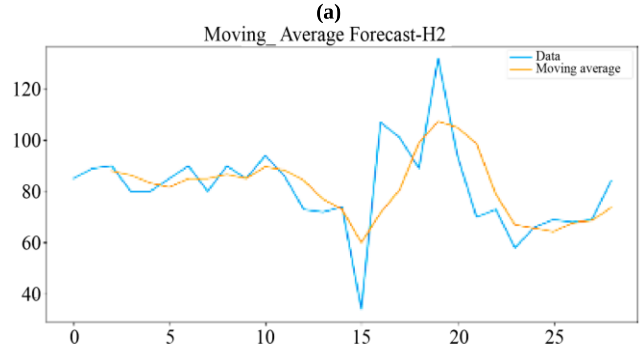
Fig. 6 Hybrid model forecast curve

Table 2. Measurement of forecast accuracy by generalized exponential smoothing

Measures (α = 0.1, β = 0.1 et γ = 1)	H2	CH4	C2H2	C2H4	C2H6	CO2
RMSE	0.733	0.289	0.008	0.297	0.325	86.834
SMAPE	0.851	0.817	136.286	0.854	0.821	0.825
MAE	0.707	0.270	0.004	0.279	0.287	81.339
MASE	0.054	0,036	0.017	0.043	0.053	0.048



(a)



(b)

Fig. 7 a) H2 forecast curve with the VARIMA model, b) H2 forecast curve for the moving average model

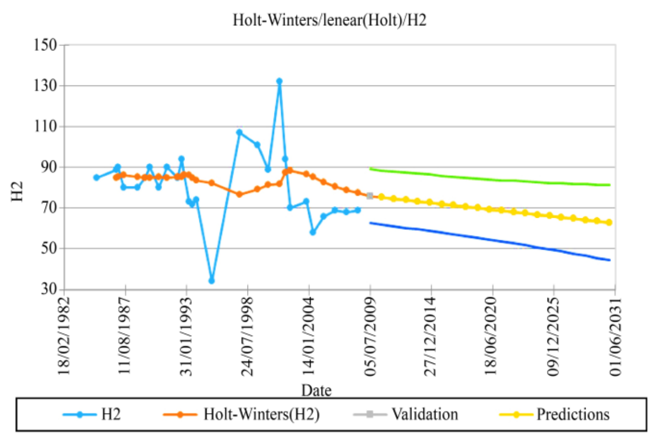


Fig. 8 Holt-Winters prediction curve with a 50% confidence interval

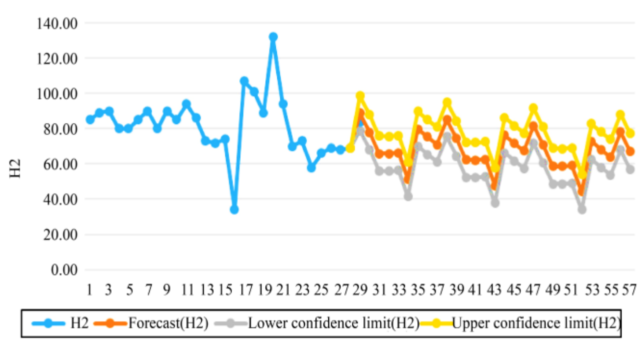


Fig. 9 Triple exponential smoothing prediction curve with 50% confidence interval

4.3.3. Proposed Model Evaluation

The various performance indicators used to evaluate the model were presented above, and their results are given in Table 2. The error values shown in Table 1 indicate that the forecast produced by the model is relatively accurate and of good quality. The low RMSE and MAE values indicate a low mean absolute error between actual and forecast values. However, the relatively high SMAPE suggests a larger relative error. The MASE of less than 1 also indicates that the model outperforms the reference model in terms of performance.

4.4. Alternative Forecasting Approach

In order to make a rigorous comparison of our model, we have decided to subject the data to other types of forecasting models. For this work, the VARIMA (......), Holt-Winter, Moving Average and triple exponential smoothing models were used, and Figures 7-9 show the results of the predictive analysis of the H2 (dihydrogen) variable.

Based on the use of VARIMA, ARIMA, Holt-Winters and MA models, Table 3 below shows the predicted values of the hydrogen(H2) variable in these models, as well as the proposed hybrid model. Table 4 shows that the hybrid model (ARIMA + GES) weighted with a coefficient of the order of [1.9 - 2.06] proposed in this work has a significantly lower set of errors than those taken individually (ARIMA, TES, Holt-Winters) and even those taken in combination (MA-GES, TES-GES, and MA-GES).

5. Conclusion

This paper proposes a predictive analysis strategy for power transformer maintenance data. It is based on the hybridization of the additive ARIMA model and generalized exponential smoothing based on the weighting of alpha, beta and gamma parameters. As maintenance in general, and that of power transformers in particular, is carried out over time and chronologically, one-hot encoding enabled us, in view of the date observations, to have a chronological date variable, as required for predictive analysis based on the arrima model.

The variables on which the study was based are of an interdependent type and were identified after a Pearson correlation matrix analysis. Of the nine dissolved gas variables, seven are interdependent, namely H2, CH4, C2H4, C2H6, CO, CO2 and N2, with correlation coefficients ranging from 0.56 to 0.89. The variables thus identified were subjected to the predictive analysis model ARIMA (2,1,1), which uses two AR terms, a differentiation of order 1 and an MA term.

This means that it captures the linear relationship between an observation and the two preceding observations, differentiates the series once to make it stationary, and also incorporates a linear dependence with the errors of the lagged observations.

Table 3. H2 variable model prediction data

Date	H2	H2-GES	H2VARIMA	H2-ARIMA	H2-MA	H2-Holt-Winters	Varima-GES	Arima-GES	MA-GES	TES-GES
1	85.00	85.85	94.686	84.985	0	0	90.268	85.419	42.925	81.94
2	89.00	89.882	74.689	87.385	0	85.000	82.286	88.633	44.941	77.901
3	90.00	90.883	87.704	88.085	88	85.440	89.294	89.484	89.441	78.346
4	80.00	80.774	75.507	85.234	86.333	85.982	78.141	83.004	83.554	73.491
5	80.00	80.766	91.445	82.531	83.333	85.409	86.106	81.648	82.049	66.073
6	85.00	85.808	83.176	81.391	81.666	84.840	84.492	83.599	83.737	82.844
7	90.00	90.85	85.416	86.953	85	84.829	88.133	88.902	87.925	83.055
8	80.00	80.74	86.267	83.575	85	85.371	83.505	82.158	82.871	75.866
9	90.00	90.83	84.489	88.752	86.666	84.806	87.662	89.793	88.751	88.017
10	85.00	85.78	87.695	82.747	85	85.348	86.736	84.262	85.388	80.113
11	94.00	94.86	85.358	93.615	89.666	85.333	90.108	94.236	92.263	78.599
12	86.00	86.77	88.891	85.204	88.333	86.306	87.831	85.987	87.552	74.505
13	73.00	73.63	87.612	83.554	84.333	86.379	80.622	78.593	78.983	68.131
14	72.00	72.62	87,856	73.783	77	85.011	80.236	73,199	74.807	60.213
15	74.00	74.63	87.774	72.681	73	83.549	81.202	73,655	73.815	75.465
16	34.00	34.22	86.786	52.059	60	82.338	60.505	43.141	47.112	52.952
17	107.00	107.95	87.076	84.877	71.666	76.765	97.512	96.413	89.807	87.679
18	101.00	101.88	85.399	77.211	80.667	79.351	93.639	89.545	91.273	91.749
19	89.00	89.75	85.322	109.894	99	81.296	87.536	99.823	94.375	80.315
20	132.00	133.17	84.249	107.677	107.333	81.923	108.711	120.425	120.253	95.967
21	94.00	94.78	83.513	101.672	105	87.288	89.147	98.227	99.891	76.721
22	70.00	70.53	82.755	96.852	98.666	88.383	76.644	83.693	84.600	64.798
23	73.00	73.56	81.719	70.266	79	86.785	77.639	71.912	76.279	58.894
24	58.00	58.40	81.269	64.756	67	85.510	69.836	61.579	62.701	65.566
25	66.00	66.48	80.359	66.634	65.666	82.586	73.419	66.556	66.073	67.289
26	69.00	69.50	79.978	63.134	64.333	80.589	74.741	66.318	66.918	66.667
27	68.00	68.49	79.496	69.922	67.666	78.976	73.993	69.205	68.077	73.264
28	69.00	69.49	79.255	68.069	68.666	77.315	74.374	68.782	69.081	68.396
29	84.00	84.64	79.188	77.052	73.666	75.836	81.913	80.845	79.153	69.909

Table 4. Errors in the different models

Measures	VARIMA-H_2 (2,1,1)	TES-H_2 (0.1,0.1,0.13)	ARIMA-H_2 (2,1,1) (0,0,0)	H-Winters (0.1, 0.1)	VARIMA-GES	ARIMA-GES	MA-GES
RMSE	17.112	18.20	11.032	23.639	8.535	5.477	12.907
SMAPE	15.803	0.17	8.939	22.566	8.359	4.510	9.844
MAE	12.612	14.35	7.220	15.384	6.279	3.549	7.075
MASE	0.967	1.17	0.554	1.180	0.482	0.272	0.543

Taking only the Mean Absolute Scaled Error (MASE) and the Mean Symmetrical Absolute Percentage Error (SMAPE), these gave 0.553 and 8.9% respectively. The MASE of less than 1 means that the selected model has good relative accuracy compared with the naive reference model. The symmetry of the percentage errors between actual and predicted values is low (8.9%), so the model provides a better forecast.

The smaller the error, the more accurate the model. To optimize the accuracy of the ARIMA model, generalized exponential smoothing with optimal coefficients $\alpha = 0.1$, $\beta = 0.1$ and $\gamma=1$ is used. The last GES coefficient is crucial in that it avoids over- or underestimation of the model parameters.

The results obtained after weighted ARIMA-GES hybridization considerably reduced the model's prediction errors, i.e. 0.27 for MASE and 4.5% for SMAPE.

This observation shows that the predictive analysis model proposed in this work is more accurate than that of ARIMA, Holt-Winters, Moving Average (MA), Triple Exponential Smoothing (TES) and many others taken individually and thus presented in Table 4. GES taken individually may give better results than hybrid models, but you have to know what you want to extract from the data. Future studies should evaluate the contribution of generalized exponential smoothing to artificial neural network forecasting models and genetic algorithms.

Abbreviations

GES: Generalized Exponential Smoothing
ARIMA: AutoRegressive Integrated Moving Average
MA: Moving Average
TES: Triple Exponential Smoothing
SES: Simple Exponential Smoothing
BES: Bulk Electric System
MOPSO: Multi-Objectif Particle Swarm Optimization
SPEA: Strength Pareto Evolutionary Algorithm
NSGA: Non-dominated Sorting Genetic Algorithm
SVDD: Support Vector Data Description
SMOTE: Synthetic Minority Oversampling Technique
mRMR: maximum Relevance Minimum Redundancy
RCM: Reliability Centered Maintenance

Acknowledgment

The authors would like to express their gratitude to American Journal Experts (AJE) for their technical support in formatting this work. Moreover, the Rio Tinto Alcan in Canada provides us with data through Pr Issouf FOFANA.

Author Contributions

Conceptualization: All authors; Data curation: M. Manyol, A. Biboum; Data visualization: M. Manyol, S. Eke, G. Olong, N.M. Matéké Max; Investigation: All authors; Methodology: All authors; Supervision: R. Mouangué, S. Eke; Writing—original draft: M. Manyol; Writing— review and editing: All authors. All authors approved the final submitted draft.

References

[1] Transformer Explosion Knocks Out Power In Northern N.J., Cbsnews, 2013. [Online]. Available: https://www.cbsnews.com/newyork/news/transformer-failure-knocks-out-power-in-northern-n-j/
[2] Daniel Kosiorowski et al., "Generalized Exponential Smoothing in Prediction of Hierarchical Time Series," *Statistics in Transition New Series*, vol. 19, no. 2, pp. 331-350, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[3] Samaher Al-Janabi et al., "Design and Evaluation of a Hybrid System for Detection and Prediction of Faults in Electrical Transformers," *International Journal of Electrical Power & Energy Systems*, vol. 67, pp. 324-335, 2015. [CrossRef] [Google Scholar] [Publisher Link]
[4] Aman Samson Mogos, Xiaodong Liang, and Chi Yung Chung, "Distribution Transformer Failure Prediction for Predictive Maintenance Using Hybrid One-Class Deep SVDD Classification and Lightning Strike Failures Data," *IEEE Transactions on Power Delivery*, vol. 38, no. 5, pp. 3250-3261, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[5] Oussama Laayati et al., "Smart Energy Management System: Oil Immersed Power Transformer Failure Prediction and Classification Techniques Based on DGA Data," *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology*, Meknes, Morocco, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[6] J.R. Jimenez-Octavio et al., "New Transformer Sizing Criteria Based on Moving Average Techniques for Railway Systems," *ASME/IEEE 2007 Joint Rail Conference and Internal Combustion Engine Division Spring Technical Conference*, Pueblo, Colorado, USA, pp. 327-332, 2009. [CrossRef] [Google Scholar] [Publisher Link]
[7] Abdolrahman Peimankar et al., "Multi-Objective Ensemble Forecasting with an Application to Power Transformers," *Applied Soft Computing*, vol. 68, pp. 233-248, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[8] Juan L. Velasquez-Contreras, Miguel A. Sanz-Bobi, and Samuel Galceran Arellano, "General Asset Management Model in the Context of an Electric Utility: Application to Power Transformers," *Electric Power Systems Research*, vol. 81, no. 11, pp. 2015-2037, 2011. [CrossRef] [Google Scholar] [Publisher Link]
[9] Allou Samé et al., "Decomposition and Classification of Functional Data for Water Consumption Analysis," *International Francophone Conference on Knowledge Extraction and Management, Clustering and Co-clustering Workshop*, Reims, France, pp. 1-11, 2016. [Google Scholar] [Publisher Link]
[10] Pere Izquierdo Gómez et al., "Data-Driven Thermal Modelling for Anomaly Detection in Electric Vehicle Charging Stations," *2022 IEEE Transportation Electrification Conference & Expo*, Anaheim, CA, USA, pp. 1005-1010, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[11] Yunus Biçen, Faruk Aras, and Hulya Kirkici, "Lifetime Estimation and Monitoring of Power Transformer Considering Annual Load Factors," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 21, no. 3, pp. 1360-1367, 2014. [CrossRef] [Google Scholar] [Publisher Link]
[12] C.C. Thompson, A.D. Stringer, and C.I. Barriga, "An Evaluation of Transformer Historical Failure Data for Facility Resiliency and Reliability," *2019 IEEE Power & Energy Society General Meeting*, Atlanta, GA, USA, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[13] Belkacem Balah, "Frontiers of Use of Monthly Rainfall Forecasting with the Arima Model in Agriculture," *1st International Seminar Agriculture 4.0: Rural Engineering at the Service of the Environment*, Algiers, Algeria, pp. 1-6, 2018. [Google Scholar] [Publisher Link]
[14] Dominique Desbois, "An Introduction to Box and Jenkins Methodology: Using ARIMA Models with SPSS," *MODULAD Review*, pp. 1-25, 2005. [Google Scholar] [Publisher Link]
[15] Régis Bourbonnais, and Virginie Terraza, *Time Series Analysis - 5th ed. Course and Corrected Exercises - Applications to Economics and Management*, Dunod, pp. 1-368, 2022. [Google Scholar] [Publisher Link]

[16] Jean-Marie Dufour, "*Trend Extraction and Seasonal Adjustment Using Moving Average Method*," McGill University, pp. 1-36, 2003. [Google Scholar] [Publisher Link]

[17] Jacques Padet, "Some Analytical Properties of Moving Averages, and their Application to the Detection of Trends in a Signal," *Entropy: Thermodynamics – Energy – Environment – Economy*, vol. 5, no. 2, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] M. Grun Rehomme, and D. Ladiray, "Centered and Non-Centered Moving Averages, Construction and Comparison," *Journal of Applied Statistics*, vol. 42, no. 3, pp. 33-61, 1994. [Google Scholar] [Publisher Link]

[19] Patrícia Ramos, and José Manuel Oliveira, "A Procedure for Identification of Appropriate State Space and ARIMA Models Based on Time-Series Cross-Validation," *Algorithms*, vol. 9, no. 4, pp. 1-14, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[20] Didier Delignieres, "*Time Series–ARIMA Models*, *EA Seminar "Sport–Performance–Health*," Mars, pp. 1-19, 2000. [Google Scholar] [Publisher Link]

[21] Youssef Hmamouche, "*Prediction of Wide Time Series*," Doctoral Dissertation, Aix Marseille University, pp. 1-108, 2018. [Google Scholar] [Publisher Link]

[22] S. Dhamodharavadhani, and R. Rathipriya, "Region-Wise Rainfall Prediction Using MapReduce-Based Exponential Smoothing Techniques," *Advances in Big Data and Cloud Computing*, *Advances in Intelligent Systems and Computing*, vol. 750, pp. 229-239, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[23] Alexei Botchkarev, "Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio," *SSRN*, pp. 1-16, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[24] Chao Chen, Jamie Twycross, and Jonathan M. Garibaldi, "A New Accuracy Measure Based on Bounded Relative Error for Time Series Forecasting," *PLoS One*, vol. 12, no. 3, pp. 1-23, 2017. [CrossRef] [Google Scholar] [Publisher Link]