A TIME SERIES ANALYSIS BASED DATA TENDENCY PREDICTION METHOD FOR DISSOLVED GAS PRODUCTION MONITORING IN POWER TRANSFORMER OIL

Zhang WEI[1], Wu RONGRONG[2], Pu JINYU[3], Deng YURONG[4]

Changing rules of time series in data of dissolved gas production produced by power transformer insulation oil monitoring are studied based on mathematical statistics. Tendency prediction model for the gas production based on autoregressive integrated moving average (ARIMA) is built. Firstly, the model takes the monitoring data sequence that has been verified by confidence testing as a random sequence, and then conducts smooth preprocessing on historical monitoring data to filter out abnormal interference prediction values. Next, model identification and analysis are conducted, and the Akaike information criterion (AIC) is used for model evaluation and model order to achieve accurate prediction for changing values of the gas production. It is validated through an actual case that the model can use historical values and current values of time series to quantify the tendency of the prediction values, and then gets confidence coefficient that is required in the tendency analysis.

Keywords: Power Transformer; Dissolved Gas in Oil; Monitoring Data; Time Series; Tendency Prediction

1. Introduction

Inner flaws of a power transformer can be found by monitoring the changes of characteristic gases (H_2、C_2H_2、C_2H_4、CH_4、C_2H_6) in transformer oil. Currently, the main technology for analyzing monitoring data is the threshold warning method which is based on static analysis. However, this method and its grade setting lack sufficient proofs [1-2]. Therefore, it is only suitable to be used in finding some dominant flaws [3-4]. If changing tendency of the monitoring data can be used to conduct the analysis dynamically, more hidden flaws can be

[1] Electric Power Science Research Institute of Guangxi Power Grid Co., Ltd, Nanning 530023, Guangxi Zhuang Autonomous Region, China. E-mail: zw_10@163.com.
[2] Electric Power Science Research Institute of Guangxi Power Grid Co., Ltd, Nanning 530023, Guangxi Zhuang Autonomous Region, China.
[3] Electric Power Science Research Institute of Guangxi Power Grid Co., Ltd, Nanning 530023, Guangxi Zhuang Autonomous Region, China.
[4] Electric Power Science Research Institute of Guangxi Power Grid Co., Ltd, Nanning 530023, Guangxi Zhuang Autonomous Region, China.

found ahead of schedule.

Currently, the most common used tendency prediction methods include support vector machine, gray model, neural network and regression model [4-9]. However, these methods cannot be used for predicting changing tendency of monitoring data of gas production in oil. The reasons include: neural network has some drawbacks such as difficulty of achieving global optimization and slow convergence speed; support vector machine is only suitable to be used in predictions that data volumes are small; gray model cannot reflect features of prediction objects and is easy to be affected by outer factors, though it needs lesser samples[10]; regression model cannot pass the hypothesis testing when the amount of samples is too large or the sampling period is too long [11-13], though it can conduct time series analysis on multiple sampling data in online monitoring applications that conduct persistence sampling analysis [14-15]. In this paper, a mathematical statistics-based method is proposed to predict the dependency relationship of the object variables in different periods of samples. At the same time, autoregressive integrated moving average (ARIMA) technology is used to construct a new tendency prediction model for time series of monitoring data of gas production in oil. The model can reveal variation feature of time series in the online monitoring data and achieve quantitative analysis for changing tendency, under the premise that abnormal prediction values are filtered.

2. Features of Monitoring Data

2.1 Time Series of Monitoring Data

Since the analysis of dissolved gas in oil can directly reflect the insulation state of a transformer, it is an important proof for fault diagnosis of electrical equipment. The advantage of transformer online monitoring is its high monitoring frequency (1 time per day). Therefore, the monitoring data contains large amounts of current and historical information of the transformer. However, a key problem of monitoring gas production in oil is that existing analysis methods are mainly based on simple threshold warning mechanisms, which cannot find hidden faults in a transformer early. Therefore, the processing time for the faults is postponed. On the other hand, in theoretical, the time series of monitoring data should have regularity and persistence, but the monitoring data in real applications often contain discrete and abnormal values that are generated by some interference factors such as equipment working conditions, mains voltage fluctuations and outer vibrations [16-17]. Therefore, the accuracy of dynamical analysis for the transformer is affected.

2.2 Smoothing Process for Historical Data

Since the interferences would affect the accuracy and the stability of the monitoring result, it is necessary to conduct smoothing preprocess on raw data sequence before the analysis. By this way, the tendency of data changes becomes quantitative. Moreover, the input parameters for the model are formed, and the abnormal values are filtered.

In the smoothing process, we assume that we have got a data sequence Y_1, Y_2, …, Y_t. Then, we use the runs test method to judge whether the sequence is an unstable sequence. If the result is "yes", we will use difference method, i.e., $Y'_{t-i} = Y_t - Y_{t-1}$, to perform tranquilization preprocess on the sequence. In each differencing, the data will be judged by the runs test. The process continues, until the data pass the stationary test and new stable sequence X_1, X_2, …, X_{t-d} are got. This process is call *d-difference.* We will take the former N groups of data (or all the data) as basis data for prediction analysis, and conduct zero-mean process (i.e., $X'_t = X_t - \bar{X}$) to get a group of new and preprocessed sequence X'_t. We are focused on the dynamical analysis of time series of the prediction values in the monitoring of dissolved gas in oil that have been preprocessed, and we use ARIMA to predict the changing tendency quantitatively.

3. Identification and Resolution of the Prediction Model

3.1 Prediction Method

ARIMA takes the data sequence of prediction objects that has passed the confidence testing as a random sequence. Then, it uses autocorrelation analysis of the time series as basis and uses mathematical model to approximately describe the sequence [18-20]. At the same time, we use ARMA model to fit the data curve of time series, when the number of the samples is larger than 50. The sampling frequency of the monitoring for dissolved gas in oil is 1 time per day, which can provide enough samples. In general, ARIMA is benefitted to solve the problems of complex model and multi-collinearity caused by the overmuch of the samples and the abnormal values, and it is more suitable to generate quantitative prediction values of changing tendency for the monitoring data. Once the model is identified, we will use Akaike information criterion (AIC) to achieve the model order determination, i.e., the monitoring data variables can be measured by the historical and current values of time series.

3.2 ARIMA Prediction Model

ARIMA prediction model firstly conducts d(d=0, 1, · · ·, n) difference processes on unstable historical monitoring data Y_t to get a new stable data sequence X_t. Then, it fits X_t as a ARMA(p,q) model. Finally, it restores and solves the original d-difference and gets the prediction data of Y_t, where ARMA (p,q)'s general expression is:

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q},\ t \in Z \tag{1}$$

In Formula (1), the former part is auto-regression; the nonnegative integer p is the order of auto-regression; φ_1, φ_2, …, φ_p are the parameters of auto-regression; X_t is correlated sequence of online monitoring data; ε_t is the stochastic disturbance sequence that obeys the independent Gaussian distribution sequence WN(0, σ^2). Moreover, the model becomes AR(p) model when q=0:

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t \qquad t \in Z \tag{2}$$

The model becomes MA(q) model when p=0:

$$X_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \qquad t \in Z \tag{3}$$

3.3 Model identification

In order to achieve the model identification, we need to calculate the autocorrelation function (ACF) $\hat{\rho}k$ and the partial ACF (PACF) $\hat{\varphi}_{kk}$ of the preprocessed data sequence X_t'. Detailed expressions are as follows:

$$\hat{\rho}_k = \frac{\sum_{t=1}^{N-k} X'_{t+k} X'_t}{N} \tag{4}$$

$$\begin{cases} \hat{\varphi}_{11} = \hat{\rho}_1 \\ \hat{\varphi}_{k+1,k+1} = (\hat{\rho}_{k+1} - \sum_{j=1}^{k} \hat{\rho}_{k+1-j} \hat{\varphi}_{kj}) \cdot \\ \qquad (1 - \sum_{j=1}^{k} \hat{\rho}_j \hat{\varphi}_{kj})^{-1} \\ \hat{\varphi}_{k+1,j} = \hat{\varphi}_{kj} - \hat{\varphi}_{k+1,k+1} \hat{\varphi}_{k,k+1-j},\ j = 1, 2, \mathrm{L}\ , k \end{cases} \tag{5}$$

According to the calculation result of above expressions and the identification rules based on ARMA (p,q) model listed in Table 1, we can make sure that X_t' conforms to the model. If the partial correlation function of the stationary sequence is truncated, and the autocorrelation function is tailed, then the sequence fits for the AR (p) model.

Table 1.

Model identification principle

Function	Identification principle		
	AR(p)	MA(q)	ARMA(p,q)
Autocorrelation function	Tailed, exponential decay or oscillation	Limited length, truncated (q steps)	Tailed, exponential decay or oscillation
Partial correlation function	Limited length, truncated (q steps)	Tailed, exponential decay or oscillation	Tailed, exponential decay or oscillation

If the partial correlation function of the stationary sequence is tailed, and the autocorrelation function is truncated, then the sequence fits for the MA (q) model. If both the partial correlation function of the stationary sequence and the autocorrelation function are tailed, then the sequence fits for the ARMA model.

3.4 Model order determination

Based on the above model identification, we will conduct parameter analysis and model order determination. In detail, we will use some methods such as sample moment estimation, least square estimation and maximum likelihood estimation to analyze the unknown parameters of ARMA (p,q), including the autoregressive coefficient, the moving average coefficient and variance of white noise, etc. Finally, the values of $\hat{\varphi}_1, \cdots, \hat{\varphi}_p, \hat{\theta}_1, \cdots, \hat{\theta}_q, \hat{\sigma}^2$ are obtained.

Then, AIC is introduced to evaluate the model and achieve the order determination. The principle of AIC is:

$$F_{\mathrm{AIC}} = N\log(\mathrm{Res}_{pq}) + 2K \tag{6}$$

where N is the total number of the sequence data; Res_{pq} is the residual sum of squares of ARMA (p,q) when its order is p and q; K=p+q.

The optimal solution will be got, if the values of p and q are taken as the model's order when AIC is the minimum. The solution is most meaningful for the quantitative analysis of the changing tendency of the monitoring data.

Hypothesis testing is needed for conducting tendency prediction analysis, i.e., whether the testing model satisfies stability and reversibility. Therefore, we require that the roots of the following formulas should be located outside the unit circles:

$$\varphi(B) = 1 - \sum_{j=1}^{p} \varphi_j B^j = 0 \tag{7}$$

$$\theta(B) = 1 - \sum_{j=1}^{p} \theta_j B^j = 0 \tag{8}$$

If the residual error sequence of the above model is not white noise, it means that there are stochastic disturbance monitoring data and we need to re-conduct model identification. By testing, we can get the prediction model of a quantitative analysis for the changing of monitoring data:

$$X'_t = \hat{\varphi}_1 X'_{t-1} + \cdots + \hat{\varphi}_p X'_{t-p} + \varepsilon_t - \hat{\theta}_1 \varepsilon_{t-1} - \cdots - \hat{\theta}_q \varepsilon_{t-q} \tag{9}$$

Based on the d-difference and standard process, we restore X'_t as the prediction result of the monitoring data Y_t. Firstly, we restore the data to the situation that has not been standardized processed, i.e., $X_t = X'_t + \bar{X}$, where

$\bar{X}$ is a constant. The process of difference restore is a recursive procedure. We set the data after d-1 differences as Z_t^{d-1}, the data after d-2 differences as Z_t^{d-2}, and so on. Then, we have:

$$X_t = Z_t^{d-1} - Z_{t-1}^{d-1},\ Z_t^{d-1} = Z_t^{d-2} - Z_{t-1}^{d-2} \cdots\cdots \tag{10}$$

And then, we have:

$$Z_t^{d-1} - Z_{t-1}^{d-1} = \hat{\varphi}_1(Z_{t-1}^{d-1} - Z_{t-2}^{d-1}) + \cdots + \hat{\varphi}_p(Z_{t-p}^{d-1} - Z_{t-p-1}^{d-1}) + \varepsilon_t - \hat{\theta}_1\varepsilon_{t-1} - \cdots - \hat{\theta}_q\varepsilon_{t-q} + \bar{X} \tag{11}$$

Let $\gamma_1^{d-1} = \hat{\varphi}_1 + 1 \cdots \gamma_p^{d-1} = \hat{\varphi}_p - \hat{\varphi}_{p-1},\ \gamma_{p+1}^{d-1} = -\hat{\varphi}_p$, the formula (11) is transformed as:

$$Z_t^{d-1} = \gamma_1^{d-1}Z_{t-1}^{d-1} + \cdots + \gamma_p^{d-1}Z_{t-p}^{d-1} + \gamma_{p+1}^{d-1}Z_{t-p-1}^{d-1} + \varepsilon_t - \hat{\theta}_1\varepsilon_{t-1} - \cdots - \hat{\theta}_q\varepsilon_{t-q} + \bar{X} \tag{12}$$

Then, we take the prediction expression of Z_t^{d-2} as

$\gamma_1^{d-2} = \gamma_1^{d-1} + 1 \cdots \gamma_{p+1}^{d-2} = \gamma_{p+1}^{d-1} - \gamma_p^{d-1},\ \gamma_{p+2}^{d-2} = -\gamma_{p+1}^{d-1}$, and we have:

$$Z_t^{d-2} = \gamma_1^{d-2}Z_{t-1}^{d-2} + \cdots + \gamma_{p+1}^{d-2}Z_{t-p-1}^{d-2} + \gamma_{p+2}^{d-2}Z_{t-p-2}^{d-2} + \varepsilon_t - \hat{\theta}_1\varepsilon_{t-1} - \cdots - \hat{\theta}_q\varepsilon_{t-q} + \bar{X} \tag{13}$$

Finally, we get the prediction equation of Y_t as follows:

$$\begin{aligned} Y_t &= \gamma_1^0 Y_{t-1} + \text{L} + \gamma_{p+d-1}^0 Y_{t-p-d+1} + \gamma_{p+d}^0 Y_{t-p-d} + \\ &\quad \varepsilon_t - \hat{\theta}_1\varepsilon_{t-1} - \text{L} - \hat{\theta}_q\varepsilon_{t-q} + \bar{X} \end{aligned} \tag{14}$$

From Equation (14) we can see that the premise of Y_t prediction is to guarantee that the number of the predicted data should be larger than $p+d$. Moreover, the effective number of the data in the prediction should be smaller than $p+d$. Therefore, ARIMA can quantitatively predict limited numbers of data values.

4. Case Study

The raw data of gas concentration collected directly from a transformer would fluctuate largely due to interference factors in the transformer, which does not benefit to the data analysis of later period. Therefore, we adopt an empirical mode decomposition technology to conduct smooth denoising on the raw data. The preprocessed data will be taken as basis in later analysis. We take a main transformer that we have monitored from 2016 as our study object. Then, we analyze the changing tendency of CH4 in its oil from the monitoring data. The process result of data filtering is shown in Fig. 1.

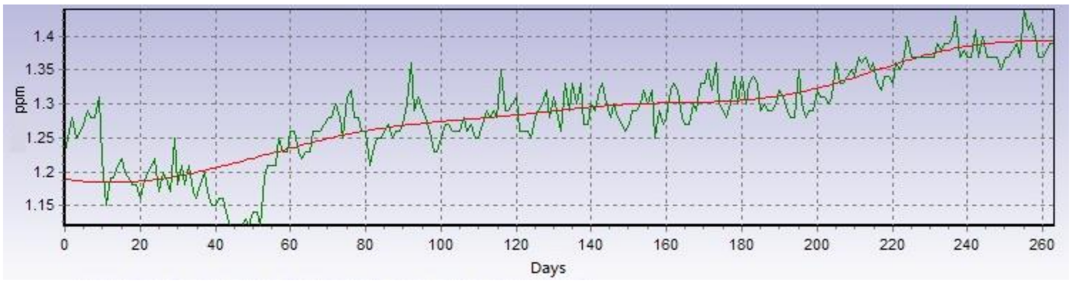

Fig. 1. Processing result of data filtering

The concentration of gas in transformer oil would increase due to the appearance of abnormal situation in the body of the equipment. In Fig. 1, the time series of CH_4's concentration have the features of long-range correlation and persistence, which show that the growing tendency of gas concentration is consistent. We use ARIMA to analyze the data changing curve of CH_4's concentration in the main transformer at the same time period, and then the result is shown in Fig. 2.

Fig. 2. Prediction and analysis results of CH_4 concentration data by ARIMA

The curve in Fig. 2 shows the data filtering curve of CH_4's concentration in the main transformer. By analysis of ARIMA prediction model, we can see that the prediction values of later 3 days are 1.43, 1.42 and 1.41. Then, a comparison between tendency prediction data and actual monitoring data is shown in Table 2.

Table 2.

Comparison of the tendency prediction data and the actual monitoring data

Data	Prediction value of concentration (μL/L)	Actual monitoring data (μL/L)	Deviation of concentration/%
264th day	1.393	1.43	2.59
265th day	1.393	1.42	1.90
266th day	1.392	1.41	1.29

We also analyzed another case of H_2. In Figure 3, the time series of H2's concentration also have the features of long-range correlation and persistence, which show that the growing tendency of gas concentration is consistent. We use

ARIMA to analyze the data changing curve of H2's concentration, and then the result is shown in Table 3.

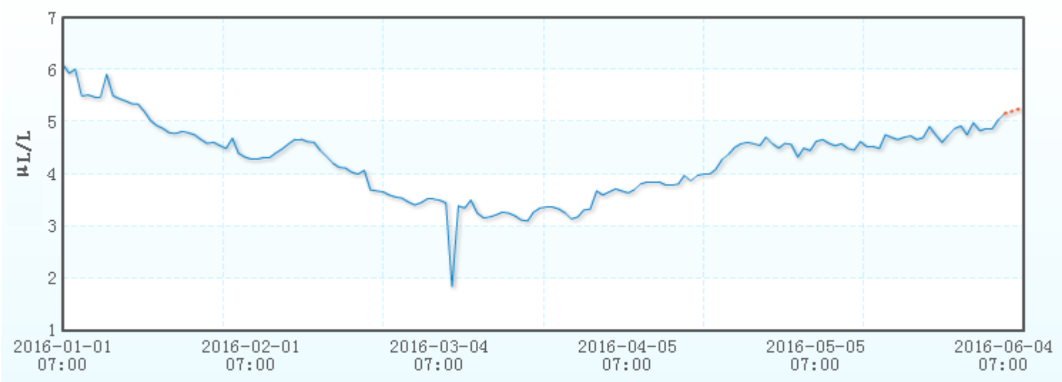

Fig. 3. Prediction and analysis results of H_2 concentration data by ARIMA

By analysis of ARIMA prediction model, we can see that the prediction values of later 3 days are 5.2, 5.24 and 5.27. Then, a comparison between tendency prediction data and actual monitoring data is shown in Table 3.

Table 3.

Comparison of the tendency prediction data and the actual monitoring data

Data	Prediction value of concentration (μL/L)	Actual monitoring data (μL/L)	Deviation of concentration/%
June 2	5.2	5.44	4.61%
June 3	5.24	5.5	4.96%
June 4	5.27	5.51	4.55%

We can see that the deviations between prediction values achieved by our model and the actual values are lower than 5%.

5. Conclusions

In this paper, a data tendency prediction method for gas production in power transformer's insulating oil is proposed. The method can distinguish the random sequence and the non-random sequence by analyzing the persistence of the monitoring data. Results of practical applications show that the method can reveal change tendency of gas in insulating oil, discover equipment potential failures and reflect equipment conditions. Though the case in the paper shows strong non-stationary in its data, the error of its prediction result is smaller than 3% comparing with actual situations. Therefore, the method is effective.

REFERENCES

[1]. *A. Akbari, A Setayeshmehr, H Borsi, et al.* Intelligent agent-based system using dissolved gas analysis to detect incipient faults in power transformers. IEEE Electrical Insulation Magazine, 26(6), 2010: 27-40.
[2]. *L. M. Krishna, J. N. Chandra Sekhar, M. Naresh, P. Samuel*. Performance analysis of grid integrated pv systems using marx multilevel inverter in different environmental conditions. Scientific bulletin, Series C: Electrical Engineering and Computer Science, Vol. 80, Iss. 2, 2018: 217-230
[3]. *Zheng Yuanbing, Chen Weigen, Li Jian, et al.* Forecasting model based on BIC and SVRM for dissolved gas oil. Electric Power Automation Equipment, 31(9) , 2011:46-49.
[4]. *Qin S J, Badgwell T A.* A survey of industrial model predictive control technology. Control Engineering Practice, 11(7) , 2003:733-764.
[5]. *J. M. Carrasco et al*., Power-Electronic Systems for the Grid Integration of Renewable Energy Sources: A Survey, in IEEE Transactions on Industrial Electronics, vol. 53, no. 4, June 2006, pp. 1002- 1016.
[6]. *Xiao Yancai, Chen Xiuhai, Zhu Hengjun*. Prediction of gas dissolved in power transformer oil by improved multivariable grey model. Power System Technology, 30(10), 2006: 86-89.
[7]. *Wahab Maa, Hamada Mm, Mohamed A.* Artificial neural network and non-linear models for prediction of transformer oil residual operating time. Electric Power Systems Research, (1), 2011: 219-227
[8]. *LI Yangliu, Zhao Xuezeng, GUO Chunzhi, et al.* Gas concentration prediction algorithm based on membrane extraction for transformer online monitoring system. Electric Power Automation Equipment, **Vol. 30**(12) , 2010:36-41.
[9]. *Gestel T, Suykens J A K, Baesaens D E, et al.* Financial time series prediction using least squares support vector machines within the evidence framework. IEEE Transactions on Neural Networks, **Vol. 12**(4), 2001:809-821.
[10]. *Wang M H.* Grey-Extension Method for Incipient Fault Forecasting of Oil-Immersed Power Transformer. Electric Power Components and Systems, **Vol. 32**(10), 2004: 959-975.
[11]. *Zhou Lijun, Wu Guangning, Zhang Xinghai, et al.* Prediction of power transformer faults based on time series of weighted fuzzy degree analysis. Automation of Electric Power Systems, **Vol.** 29(13) 2005:53-55.
[12]. *Sonnenburg S, Ratsch G.* Large scale multiple kernel learing. Journal of Machine Learning Research, **Vol.7**, 2006:1-18.
[13]. *CHEN Jindong, PAN Feng*. Online support vector regression predictive control algorithm based on particle swarm optimization. Information and Control, 42(6) , 2013: 723-728,734.
[14]. *Tang Junjie, NIU Huanna, Yang Minghao*. Periodic autoregressive short-term forecasting method based on the linear correlation analysis. Power System Protection and Control, 38(14), 2010:128-133.
[15]. *Laurentiu Duca, Cornel Popescu*. Improving PSO based algorithms with the domain-shrinking technique for electromagnetic devices optimization. SCIENTIFIC BULLETIN, Series C: Electrical Engineering and Computer Science, Vol. 80, Iss. 1, 2018: 179-190.
[16]. *Lin C H, Wu C H, Huang P Z.* Grey clustering analysis for incipient fault diagnosis in oil-immersed transformers. Expert Systems with Applications, 16(10) , 2008: 1-9.
[17]. *HU Jinsong, BAO Jilong, ZHOU Fangjie, et al.* Model matching-based transformer chromatogram peaks identification algorithm. Automation of Electric Power Systems, 2005, 29(21):85-87.
[18]. *Yadav R K, Balakrishnan M.* Comparative evaluation of ARIMA and ANFIS for modeling of wireless network traffic time series. EURASIP Journal on Wireless Communications and

Networking, (1), 2014:15.

[19]. *Laner M, Svoboda P, Rupp M.* Parsimonious fitting of long-range dependent network traffic using ARMA models. IEEE Communications Letters, 17(12), 2013: 2368-2371.

[20]. *Wang P, Billinton, R.* Time-sequential simulation technique for rural distribution system reliability cost/worth evaluation including wind generation as alternative supply. IEEE Proceedings: Generation, Transmission and Distribution, **Vol. 148** (4), 2001:355-360.