# A Condition Monitoring System for Wind Turbine Generator Temperature by Applying Multiple Linear Regression Model

Khaled B. Abdusamad, David Wenzhong Gao
Department of Electrical and Computer Engineering
The University of Denver
Denver, Colorado, United State
khaledabdusamad@yahoo.co.uk, Wenzhong.Gao@du.edu

Eduard Muljadi
National Renewable Energy Laboratory
15013 Denver West Parkway
Golden CO 80401-3393
eduard.muljadi@nrel.gov

***Abstract*—The development and implementation of condition monitoring system become very important for wind industry with the increasing number of failures in wind turbine generators due to over temperature especially in offshore wind turbines where higher maintenance costs than onshore wind farms have to be paid due to their farthest locations. Monitoring the wind generators temperatures is significant and plays a remarkable role in an effective condition monitoring system. Moreover, they can be easily measured and recorded automatically by the Supervisory Control and Data Acquisition (SCADA) which gives more clarification about their behavior trend. An unexpected increase in component temperature may indicate overload, poor lubrication, or possibly ineffective passive or active cooling. Many techniques are used to reliably predict generator's temperatures to avoid occurrence of failures in wind turbine generators. Multiple Linear Regression Model (MLRM) is a model that can be used to construct the normal operating model for the wind turbine generator temperature and then at each time step the model is used to predict the generator temperature by measuring the correlation between the observed values and the predicted values of criterion variables. Then standard errors of the estimate can be found. The standard error of the estimate indicates how close the actual observations fall to the predicted values on the regression line. In this paper, a new condition-monitoring method based on applying Multiple Linear Regression Model for a wind turbine generator is proposed. The technique is used to construct the normal behavior model of an electrical generator temperatures based on the historical generator temperatures data. Case study built on a data collected from actual measurements demonstrates the adequacy of the proposed model.**

## I. INTRODUCTION

Although the number of wind turbine failures due to excessive generator temperatures has been increasing,wind energy development is still a very significant trend in the coming years since it is one of the most effective types of renewable energy. Condition monitoring system (CMS) plays a pivotal role for condition-based maintenance and repair, which can be more beneficial than corrective and preventive maintenance [1,2,3,4].To achieve this objective, there is a need to develop an active fault prediction algorithms which shall be the basis of CMS.Temperature is extensively monitored in wind generators; for example, temperature sensing is used to monitor specific areas of the stator core and the cooling fluids of large electrical machines like wind turbine generators. Such measurements can only give indications of overall changes taking place within the machine but they are extremely effective if mounted and monitored in carefully selected sites. Generator temperatures have direct relation with the electrical loads and ambient conditions; consequently, when temperature measurement is combined with information of the system conditions, an effective condition monitoring can be achieved. Many mathematical methods are used to construct the normal operating model for wind turbine generator temperature and then at each time step the model is used to predict the generator temperature. For example, a temperature trend analysis method based on the nonlinear state estimate technique (NSET) is proposed in which the differences between predication and actual values are used as an important indication to study the potential fault that may occur in wind turbine generator based on SCADA. The proposed technique is used to construct the normal behavior model of the electrical generator temperature, and the technique can be utilized to identify dangerous generator over temperature before damage that could occur and cause shut down of the wind turbine [5]. Another paper investigates that mechanical characters can be used to diagnose the faults that can happen in generator by simulating the impacts of wind turbine components working under different conditions. The authors proposed a method using the possibility of detecting mechanical and electrical faults in wind turbines by applying wavelet transform through analyzing the power signal correctly using a valid signal processing technique [6]. They assumed in their work that when the applied torque varies slowly relative to the electrical grid frequency, a quasi-steady state approach (when the mechanical torque is approximately equal to electric torque) may be taken for the analysis. Condition monitoring of wind generator was discussed in another paper by using the time and frequency domain analysis [7]. The authors emphasize that by monitoring the stator and rotor line current trend when both stator and generator rotor are under unbalanced force, the detección of generator faults is feasible. This paper considers the Multiple Linear Regression analysis as one of the most widely used statistical techniques for analyzing multifactor data and is used for investigating and modeling the relationship between variables. The variables must be logical and selected carefully to achieve reliable results. Therefore, Multiple Linear Regression Model (MLRM) can be the basis for a new approach to predict and monitor the temperatures inside the wind turbine generators efficiently and reliably,

by computing the correlation between the observed values and the predicted values of the criterion variable based on the historical generator temperatures. The arrangement of this paper pursues the following steps. Section II presents knowledge and specifications about a wind turbine, generator, cooling system, and the available SCADA data which are used as a case study to test the proposed model in this paper. Section III explains how the MLRM as a technique is constructed and then used to predict the generator temperatures. Section IV is concentrated on the selected variables that are used in order to construct the MLR model of generator temperature. Section V presents a case study used to test the capability of MLRM technique to predict the generators temperatures and detect early fault. The obtained results and analysis are presented in section VI. Section VII provides discussion, conclusions and suggestions for further research.

**Nomenclature:**

| | |
|---|---|
| GT | Generator temperature |
| GP | Generator power |
| OT | Ambient or outside temperature |
| NT | Nacelle temperature |
| CT | Generator cooling temperature |
| y | Dependent variable (experimental value). |
| $\hat{y}$ | The predicted dependent variable in the model |
| $\bar{y}$ | The experimental value mean (mean value of y) |
| k | Number of independent variables |
| $X_i\ (i = 1,2,\dots,k)$ | The i[th] independent variable from total set of k variables |
| $\beta_i\ (i = 1,2,\dots,k)$ | The i[th] coefficient corresponding to $X_i$ |
| $\beta_0$ | The intercept coefficient (or constant) |
| i=1,2,3,….,k | Independent variables' index |
| N | Number of observations ( experimental data points) |
| $\varepsilon$ | Residual (the difference between the experimental and predicted value) |
| $SS_{RES}$ | The residual sum square |
| $SS_R$ | The regression sum square |
| $SS_T$ | The total sum of squares |
| $MS_{RES}$ | The residual mean sum square |
| $MS_R$ | The regression mean sum square |
| $\alpha$ | The confidence interval percent |
| $\sigma^2$ | The error variance of term y |
| $\delta^2$ | The residual mean square |
| $C_{jj}$ | The j[th] diagonal element of the $(X'\,X)^{-1}$ matrix |
| $F_0$ | The significance of regression statistical value. |
| T | Statistical value (the ratio of the coefficient to its standard error) |
| $R^2$ | The coefficient of determination. |
| $R^2_{Adj}$ | The adjusted coefficient of determination. |
| $S_{ii}, i = 1,2,,,.,k$ | The corrected sum of squares for regressor $X_i$ |
| $S_{jj}, j = 1,2,,,.,n$ | The corrected sum of squares for regressor $X_j$ |
| $r_{ij}$ | The correlation between the regressor $X_i$ and $X_j$ |
| $\hat{b}_i$, i=1,2,….,k | The standardized regression coefficients. |
| $w_i, i = 1,2,\dots,k$ | The new length standardized regression scaling (the independent variables importance in the model) |
| $SS_{LOF}$ | The lack-of-fit sum of squares. |
| $SS_{PE}$ | The pure-error sum of squares |
| $F_{LOF}$ | The Lack-of-fit statistical value |
| $df_{LOF}$ | The lack-of-fit degree of freedom. |
| $df_{PE}$ | The pure error degree of freedom. |
| $h_{ij}$ | The $ij$[th] element of the hat matrix H. |
| VIF | The variance inflation factor. |
| Z | The reduced variable. |
| $\bar{X}_{i_{old}}$, i= 1,2,…,k | The mean value of the old data |

## II. THE SELECTED GENERATOR AND SCADA PARAMETERS INFORMATION

The data were collected from a variable speed wind turbine with rated power of 450 KW and rated speed 17m/s. The generator used is three phase permanent magnetic type 440/660 V 60 Hz with speed of 1500rpm and is forced air-cooled using a closed-loop with air to air heat exchanger to discharge heat to the surrounding. The cut-in speed and cut-out wind speed of the turbine is 4.5m/s, and 24m/s respectively [8]. More than 80 parameters are measured and available every 10 minutes by SCADA system. The data behavior is analyzed, then the obtained data are inserted into the proposed model to apply the condition monitoring accurately. The collected data are recorded in 23/05/2011 and covers the period from 05:50 pm to 06:00 pm which means 60,000 samples obtainable through 600 second. The covered period is enough to obtain information about the generator performance and predict early faults that can occur due to increase in the generator temperature. Each record includes much information about the selected wind turbine like wind speed, output power, stator current, stator voltage, ambient temperatures, nacelle temperatures, generator stator winding and cooling air temperature,…,etc. For simplicity, we consider the generator stator winding temperature as the generator temperature. Commonly, generators have specific allowable temperature limitation and according to the manufacturers fault handbook of the selected wind turbine, an over temperature alarm will occur and the wind turbine shuts down when the generator temperature reaches the limitation temperature (140C°) through a continuous duration of 60 seconds, and when the temperature falls to below 130C°, the wind turbine will restart [8]. The obtained historical data can be analyzed by using the proposed model and the model output can be compared with the actual measured data to compute the residual mean value and investigate the model shape, which leads to determine possible faults due to increase in the generator temperature.

## III. MLRM CONSTRUCTION FOR GENERATOR TEMPERATURE

Multiple linear regression model is one of the most popular statistical techniques that uses more than one regressor (independent variable) to predict the behavior of a response (dependent variable) by modeling and investigating the relationship between those variables [9]. The technique is described as follows:

### *A. Estimate the Model Parameters*

Let there be $k$ independent variables of interest in a process, $x_1, x_2, x_3, \dots\dots, x_k$, and suppose $y$ is a response and dependent variable to the variables $x$. The multiple regression model that might describe the relation between the dependent and independent variables to predict the outcome of the response y in future can be defined as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots\dots\dots + \beta_k X_k + \varepsilon$$

In general the response y may be related to k regressor or predictor variables, and the regression coefficients $\beta_0, \beta_1, \beta_2, \dots\dots, \beta_n$ can be estimated by using method of least squares [9,10].

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ : \\ : \\ \beta_k \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_{11} & X_{12} \dots \dots . X_{1K} \\ 1 & X_{21} & X_{22} \dots . . X_{2K} \\ : & : & : \quad : \\ 1 & X_{n1} & X_{n2} \dots . \dots . X_{nk} \end{pmatrix},$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ : \\ : \\ y_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ : \\ : \\ \varepsilon_n \end{pmatrix}, \quad \beta = (X' X)^{-1} X' y$$

Then the predicted dependent value ($\hat{y}$) can be computed as:

$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \dots \dots . + \beta_k X_k$ [9,10].

*B. Find the Residuals Between the Observed Values and the Corresponding Fitted Valued*

The residuals $\varepsilon$ (the difference between the observed values of $y$ and the corresponding predicted values $\hat{y}$) play an important role in evaluation of the adequacy of the fitted regression model and the shape of the model. Moreover; by determining and analyzing the relation between the residuals ε and corresponding fitted values $\hat{y}$, the model deficiencies show up clearly [10,11]. The error variance of term $y$ is $\sigma^2$ which can be determined by using the following equation:

$$\sigma^2 = \frac{SS_{Res}}{n-p}$$

The residual mean square and residual sum square can be calculated respectively as follows:

$$MS_{Res} = \frac{SS_{Res}}{n-k}, \quad SS_{Res} = y' y - \beta' X' y$$

Where $n-p$ =Residual degree of freedom, and $p = k-1$ where $k$ is the regression degree of freedom. The total sum of squares $SS_t$ is partitioned into a sum of squares due to regression $SS_R$ and a residual sum of squares $SS_{Res}$. Thus,

$SS_T = SS_R + SS_{Res}$, $SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, and

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

The regression mean square $MS_R = \frac{SS_R}{k}$ , [9,10].

*C. Find the Coefficients' Confidence Intervals*

The next step is finding the coefficients' confidence intervals of the proposed multiple regression model by utilizing the next formula:

$$\beta_i - \left[t_{\frac{1-\alpha}{2},n-p}\right] * \sqrt{\delta^2 * C_{jj}} \leq \beta_{jj} \leq \beta_i + \left[t_{\frac{1-\alpha}{2},n-p}\right] * \sqrt{\delta^2 * C_{jj}}$$

where $\delta^2$ is the residual mean square, $\alpha$ is the confidence interval percent ($\alpha = 95\%$ in the model assumption), and $C_{jj}$ the j[th] diagonal element of the $(X' X)^{-1}$ matrix [9,10].

*D. Measure the Model Adequacy and Linearization*

To measure the normality (The residual points behavior) and linearization (whether a linear relationship exists between the response variable and regressor variables) of the proposed model, certain statistic tests of hypotheses about the model parameters are useful in measuring model adequacy.

*1) Test Significance of Regression:* Test of significance of regression is needed to investigate whether a linear relationship between the response y and any of the regressor variables is present. The statistical concept of this test emphasizes that at least one of the independent variables $X_1, X_2, \dots X_k$ is related strongly to the model [9,10]. The hypothesis of this test is as follows: (If the significance of regression statistical value ($F_0$) is more than proposed F value ($F_{TABLE}$), then the hypothesis of $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ is rejected). Which means:

$$F_0 \text{ should be} > F_{1-\alpha,k,n-k-1}, \; F_0 = \frac{MS_R}{MS_{Res}}$$

*2) The Coefficient of Multiple Determination:* The coefficient of determination $R^2$ and the adjusted $R^2$, denoted $R^2_{Adj}$ are measures to test the goodness of fit of the proposed model. They can be calculated as follows:

$$R^2 = \frac{SS_R}{SS_T}, \qquad R^2_{Adj} = 1 - \frac{\frac{SS_{Res}}{n-p}}{\frac{SS_T}{n-1}}$$

The high value of $R^2$ or $R^2_{Adj}$ does not necessarily denote that the regression model is suitable. In many cases adding a new independent variable to the model may causes worse results when the error mean square for the new model ($MS_{RES_{NEW}}$) is larger than the error mean square of the older model ($MS_{RES_{OLD}}$), although the new model will show an increased value of $R^2$ or $R^2_{Adj}$.

*3) Residual Analysis:* Previously residuals had been defined as the difference between the observation and fitted value of the dependent variable. By computing the residual, the deviation between the data and the regression model can be viewed. Therefore; residual cam be considered as a measure of the variability in the response variable not explained by the regression model. It is also convenient to think of the residuals as the realized or observed values of the model errors. Thus, any departures from the assumption on the errors should show up in the residuals. Analysis of the residuals is an effective way to discover several types of model inadequacies. Plotting residuals is a very effective way to investigate how well the regression model fits the data. The normal probability plot is a graphical tool for comparing a data set with the normal distribution, and can be considered as a method of checking the normality assumption of the proposed model. Figure 1 displays an acceptable normal probability plot in which the points lie approximately along a straight line [9, 10].

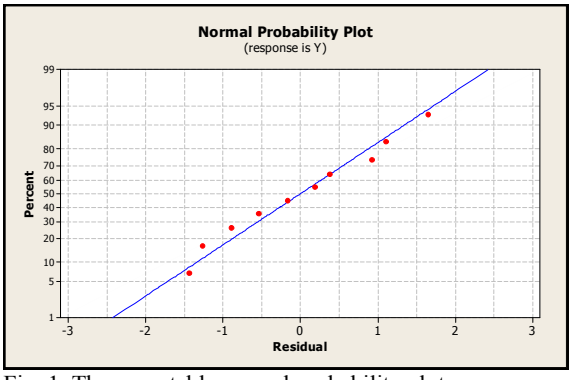


Fig. 1. The acceptable normal probability plot

Plot of residual against the fitted values $\hat{y}_i$ is also useful for detecting several common types of model inadequacies. The ideal model should have residuals contained in a horizontal band (the points on the plot show no pattern or trend) as in Figure 2 which indicates that there are not obvious defects and there is no sense in adding new independent variable to the model [9, 10].

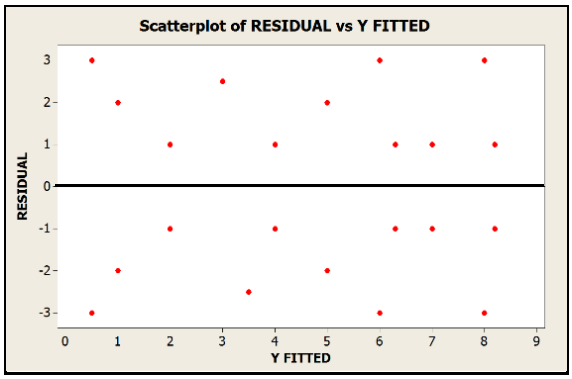

Fig. 2. Scatter plot of acceptable fitted values

*4) PRESS Statistic Test:* The predicted residual sum of squares (PRESS) is a measure of regression model validity and potential performance in prediction, and it can be defined as the sum of the squared residuals. When an observation falls outside the general trend of the data, it considers an influential observation and adversely affects the model. The presence of influential observation can be exposed by computing the PRESS statistic value. Therefore, PRESS statistic is considered as a measure of how well a regression model will be in predicting new data. A model with a small value of PRESS is desired and can be compared with the PRESS residual and computed as follows:

$$PRESS\ \ statistic\ value = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_{ij}}\right)^2$$

where $h_{ij}$ is the leverage for the $ij_{\text{th}}$ element of the hat matrix (H), H= $X\ (X'X)^{-1}X'$

The desired model should have smaller PRESS statistic value than the residual error value of the model where residual error is $\sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2$ [9, 10].

*5) Apply a formal Test for Lack of Fit:* This is a test to confirm if there is a linear relationship between the dependent response y and any of the regressor variables $X_1, X_2, X_3, \ldots X_K$. It can consider this test as an overall test of model adequacy. The requirements of the formal statistical test for the lack of fit of a regression model are the normality, independence, and constant-variance requirements to confirm that the tentative model adequately described the data [9,10]. The lack-of-fit sum of squares is found by the following:

$$SS_{LOF} = SS_{RES} - SS_{PE}$$

where $SS_{PE}$ is the pure-error sum of squares which can be calculated as

$$SS_{PE} = \sum_{i=1}^{m} \sum_{j=1}^{ni} (y_{ij} - \bar{y}_i)^2,$$

where $\bar{y}_i$ is the average of the $n_i$ observations at $X_i$, $j$ the number of measurements, m the degree of freedom. The hypothesis that says the model adequately described the data when Lack-of-fit F test $F_{LOF} > F_{\alpha, df_{LOF}, df_{PE}}$ is rejected, where $F_{LOF}$ can be calculated from the next relation:

$$F_{LOF} = \frac{mean\ square\ of\ lack\ of\ fit}{mean\ square\ of\ pure\ erroe} = \frac{SS_{LOF}/df_{LOF}}{SS_{PE}/df_{PE}}$$

*E. Transformation to Linearize the Model*

When the model fails to exceed the statistical hypothesis tests, there is a required transformation on the regressor variables $(X_1, X_2, X_3, \ldots X_k)$ since the relationship between the dependent variable (y) and one or more of the regressor variables is nonlinear even though the condition of normal distribution is satisfied (the residual is almost normally distributed). Therefore, a nonlinear function can be linearized by using a proper transformation [9, 10]. There are many linearized functions that can be used for this purpose. Polynomial regression models are widely used in situation where the response is curvilinear. For example, a second-order polynomial model with two variables would be:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$$

To linearize the proposed model, several steps should be taken as follows:

1) *Apply the Mean-Centering Method:* One of the most common transformation methods is mean-centering technique which calculates the mean of each independent variable and compute new value for each independent variable by subtracting the old independent variable value at each observation from its mean. Mean centering method works perfectly when polynomial regression technique is applied and used to leave the multicollinearity problem, which occurs when one or more of the regressor variables are strongly correlated together [9,10]. Therefore, the new regressor data will be obtained by applying the following formula for all the independent variables:

$$X_{i_{NEW}} = X_{i_{old}} - \bar{X}_{i_{old}}$$

2) *Apply Multicollinearity Test:* Multicollinearity is a problem in multiple regression that occurs when one or more of the regressor variables are robustly correlated with each other, which is undesirable in the proposed technique. Therefore, high multicollinearity between the independent variables results in large variance and covariance for the least-squares estimators of the regression coefficients which causes highly different estimates of the model parameters and leads the coefficients insignificant (the coefficients are unstable, and their standard errors are large). A very simple measure of multicollinearity is calculating the variance inflation factors (VIF) which can be obtained from the following formula:

$$VIF_j = (1 - R_i^2)^{-1}$$

where $R_i^2$ is the coefficient of determination obtained when a particular independent variable is regressed with degree of freedom equals the total number of variables (y and Xs) $-1$. Practical experience emphasizes that when the variance inflation factors values VIFs exceeds 5, the correlated regression coefficients are poorly estimated because of

multicollinearity. SPSS and Minitab statistical softwares [11, 12] automatically perform a tolerance analysis and will not adopt the model results with tolerance < 0.2 for each variable inserted into the regression model.

$$Tolerance = \frac{1}{VIF}$$

3) *Standardized Regression Coefficients:* In order to determine which independent variable is the most important to compute the response y value (dependent variable), length scaling method can be used for the independent variables and response variable since the dimension of the dependent variable and some of independent variables are different and the units of the regression coefficients are units of the dependent variable y/units of independent variable $x_i$. For this reason, it is beneficial to act with scaled regressor and response variables by creating dimensionless regression coefficients. The corrected sum of squares for regressor $X_i$:

$$S_{ii} = \sum_{i=1}^{n}(X_{ij} - \bar{X}_i)^2,$$

where i = 1,2,.., k, and j= 1,2,..,n. The simple correlation between the $X_i$ and $X_j$ is $r_{ij}$:

$$r_{ij} = \frac{S_{ij}}{(S_{ii}S_{jj})^{1/2}}$$

In this scaling, each new regressor is $w_i$ [9, 10].The correlation matrix $w'w$ and the standardized regression coefficients matrix $\hat{b}$ can be calculated as follows:

$$w'w = \begin{pmatrix} 1 & r_{12} & r_{13} \dots\dots & r_{1K} \\ r_{12} & 1 & X_{23} \dots\dots & X_{2K} \\ r_{13} & r_{23} & 1 \dots\dots & X_{3K} \\ : & : & : & : \\ r_{1k} & r_{2k} & r_{3k} \dots\dots & 1 \end{pmatrix}, \quad \hat{b} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ : \\ : \\ \hat{b}_k \end{bmatrix} = (w'w)^{-1} \; r_{iy}$$

By determining the values of the standardized regression coefficients, the most significant regressor in the proposed model will be discovered. The previous steps can be summarized in the next flowchart:

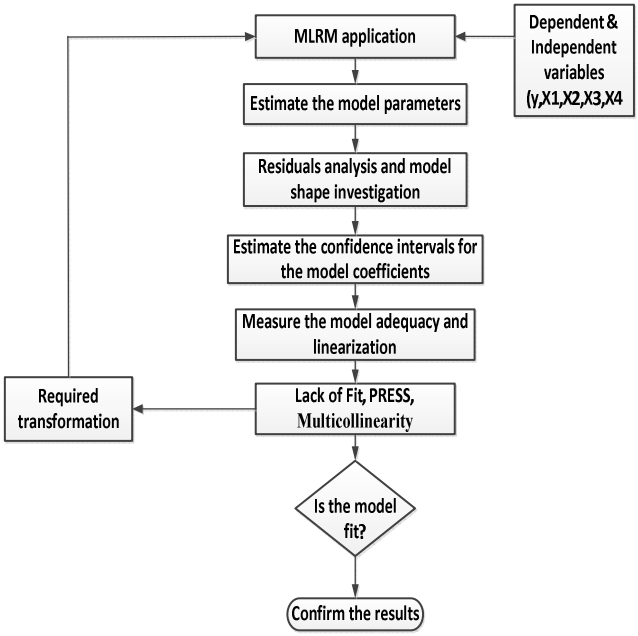


Fig. 3. The multiple regression technique.

## IV. THE SELECTED VARIABLES OF THE MODEL

The selected variables that are related to the proposed model can be defined as follows:

1) *Generator Power (GP):* Generator power has direct effect on the generator temperature. The stator current in the generator will be increased when the electrical load is high which leads to increase of the generator output power, and temperature of generator.

2) *Ambient or Outside Temperature (OT):* The significant and frequent rise in the outside temperature leads to increase of the generator temperature.

3) *Nacelle Temperature (NT):* The nacelle temperature has close relevance with the generator temperature since the generator itself is located inside the nacelle component.

4) *The Cooling Air Temperature of the Generator (CT):* The cooling air temperature of the generator has strong relationship with the generator stator cooling condition, which affects its temperature directly

5) *The Generator Stator Winding Temperature (GT):* represents the dependent variable, and is dependent on the previous independent variables. Moreover, all variables data are available to predict the generator temperatures to detect the potential faults, consequently; protect wind turbines from damage and decrease the maintenance and operation cost [5,6,13,14,15].

## V. CASE STUDY

As already mentioned, the collected data by SCADA system provides enough knowledge for effective monitoring. According to the manufacturers fault handbook (manual) of the selected wind turbine, the studied wind turbine shuts down when the generator temperature amounts to 145 Cº over a continuous period of 60 seconds, and restarts when the generator temperature drops to 130 Cº [8]. The considered independent parameters in the proposed model are the generator power, outside temperature, nacelle temperature, and generator cooling air temperature. The dependent variable in the model represents the generator stator temperature. According to the obtained data, Figurer 4 presents the selected independent variables behavior over a continuous period of 600 seconds.

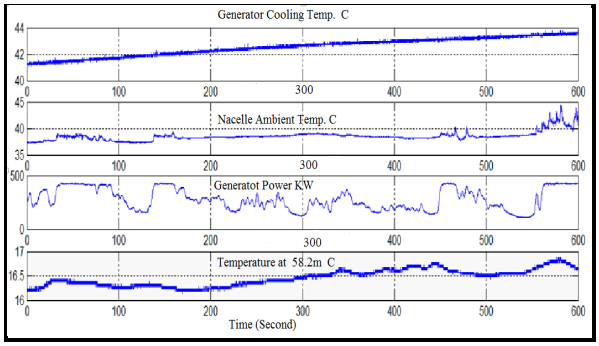


Fig. 4. Selected independent variables behavior over the time.

By using Minitab or SPSS statistical softwares and inserting the related data to the model, model coefficients and regression sum of squares of the temporary output

regression model give the first indication of the model. The initial regression equation is:

$$GT = 110 + 0.0103\,GP + 0.049\,OT + 0.0079\,NT - 0.006\,CT$$

It can be seen that the model coefficients lay within the model output confidence intervals at 95% confidence interval.

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| **Source** | ***df*** | ***SS*** | ***MS*** | ***F*** | ***P*** |
| ***Regression*** | 4 | 61566 | 15392 | 12599223.64 | 0.000 |
| ***Residual Error*** | 59995 | 73 | 0 | | |
| ***Lack of Fit*** | 25094 | 68 | 0 | 16.83 | 0.000 |
| ***Pure Error*** | 34901 | 6 | 0 | | |
| ***Total*** | 59999 | 61640 | | | |

PRESS = 73.3045 R-Sq(pred) = 99.88% R-Sq = 99.9%

Moreover, the coefficient of determination $R^2$ and the adjusted $R^2$ are very high (99%) which indicates that adding a new term may make the regression model worse if the error mean square for the new model $MS_{Res_{New}}$(when adding new independent variable to the model) is larger than the error mean square of the older model $MS_{Res_{old}}$ (without adding new independent variable to the model).

### *A. Test Significance of Regression*

The software results confirm that the generator temperature is related to all selected regressor variables since the statistical $F_0 = 12599223.64 > F_{\alpha,df_{REG},df_{RES}} = 2.37$ where $\alpha$ is the percent confidence interval which equals to 95% in this present work. $F_{\alpha,df_{REG},df_{RES}}$ is collected from the F statistical distribution tables [10,11].

### *B. PRESS Statistic and Lack of Fit Tests Results*

When applying the PRESS statistic and lack of fit test to find the model adequacy, it becomes clear that the model failed to pass these tests. It can be seen when computing PRESS statistic which is equal to 73.3045 and is bigger than the residual sum of square which is equal to 73. Moreover, there is a lack of fit in the model and the regression function is not linear since the lack-of-fit test statistic $F_{LOF} = 16.83 > F_{\alpha,df_{LOF},df_{PE}} = 1$. The reason why the model does not fit the data is that the relationship between the generator temperature and some of the regressor variables (ambient, nacelle, and cooling air temperatures) is nonlinear. Therefore, appropriate transformation on the regressor variables and the dependent variable is necessary to let the model exceed the proposed statistical test.

### *C. Polynomial Regression Model*

Polynomial regression models are widely used in situations where the response is curvilinear to provide better results. By plotting the relationships between the observed generator temperatures and all the selected variables, it is found that the quadratic curve is very suitable for the majority of variables. This means that the proper selection of the fitting model for the four independent variables is the polynomial regression model. Since there are four independent variables, the fitting polynomial regression model of fourth-order response surface in four variables is as follows:

$$GT = \beta_0 + \beta_1.GP + \beta_2.OT + \beta_3.NT + \beta_4.CT + \beta_5.GP^2 + \beta_6.OT^2 + \beta_7.NT^2 + \beta_8.OT^2 + \beta_9.GP^3 + \beta_{10}.OT^3 + \beta_{11}.NT^3 + \beta_{12}.CT^3 + \beta_{13}.GP^4 + \beta_{14}.OT^4 + \beta_{15}.NT^4 + \beta_{16}.CT^4 + \beta_{17}.GP.OT + \beta_{18}.GP.NT + \beta_{19}.GP.CT + \beta_{20}.OT.NT + \beta_{21}.OT.CT + \beta_{22}.NT.CT + \beta_{23}.GP.OT.NT + \beta_{24}.GP.OT.CT + \beta_{25}.GP.NT.CT + \beta_{26}.OT.NT.CT + \beta_{27}.GP.OT.NT.CT \quad [10, 11].$$

The obtained results are given in the following table:

| **Source** | **DF** | **SS** | **MS** | **F** | **P** |
|---|---|---|---|---|---|
| ***Regression*** | 27 | 61615.6 | 2282.1 | 5657660.45 | 0.000 |
| ***Residual Error*** | 59972 | 24.2 | 0.0 | | |
| ***Lack of Fit*** | 25071 | 18.6 | 0.0 | 4.63 | 0.000 |
| ***Pure Error*** | 34901 | 5.6 | 0.0 | | |
| ***Total*** | 59999 | 61639.8 | | | |

PRESS = 24.2159 R-Sq(pred) = 99.96% R-Sq = 99.9%

From the previous results it becomes clear that there is improvement in the new model. The statistic value $F_{LOF}$ is reduced to 4.63 which implies that the polynomial regression model in four regressor variables is a proper selection. However, The PRESS statistic value = 24.2159 > Residual sum of square = 24.2 which means that some observations in the model falls outside the general trend of the data and affect the model quality. By applying the mean-centering method, this problem can be solved. In mean-centering method each dependent variable is subtracted from its mean and the new values of each variable is inserted into the polynomial regression model. The new obtained result is improved since the model passed the lack of fit and PRESS statistic tests. The output results confirm that the regressor variables are nearly perfectly linearly related, and in such situations the inferences based on the regression model can be misleading (deceptive) since the variance inflation factors (VIF) for all variables are very high ($> 5$) [9, 10]. The obtained correlation matrix strongly confirms that the variance inflation factors for all variables exceed 5 which means that the problem of multicollinearity is existing. Table 1 and 2 display the variance inflation factors and correlation of some independent variables, and we can see from the correlation matrix that the correlation values between some of independent variables inserted to the polynomial regression model are very high which cause problem of multicollinearity in the proposed model and thus the model is not proper.

| **Source** | **DF** | **SS** | **MS** | **F** | **P** |
|---|---|---|---|---|---|
| ***Regression*** | 5 | 61600 | 12320 | 18508202.24 | 0.000 |
| ***Res. Error*** | 59972 | 40 | 0.0 | | |
| ***Lack of Fit*** | 22328 | 16 | 0.0 | 0.984 | 0.000 |
| ***Pure Error*** | 37666 | 24 | 0.0 | | |
| ***Total*** | 59999 | 61640 | | | |

PRESS = 39.9442 R-Sq(pred) = 99.94% R-Sq = 99.9%

Table I

COLLINEARITY STATISTICS

| **Model** | **Tolerance** | **VIF** |
|---|---|---|
| **CT** | 0.013 | 77.330 |
| ***NT²*** | 0.003 | 359.768 |
| ***NT³*** | 0.001 | 878.063 |
| ***NT⁴*** | 0.004 | 258.316 |
| ***OTNT*** | 0.005 | 217.610 |
| ***GPOTNT*** | 0.005 | 202.966 |
| ***GPOTNTCT*** | 0.008 | 126.139 |

Table II

SAMPLE OF CORRELATION MATRIX

| *MODEL* | $NT^3$ | $OT^4$ | $NT^4$ | NT | GPNT | OTNTCT |
|---|---|---|---|---|---|---|
| *GPOTNT* | 0.823 | 0.801 | 0.716 | 0.83 | 0.854 | 0.927 |
| *GPNTCT* | 0.807 | 0.7 | 0.7 | 0.83 | 0.83 | 0.884 |
| *GPNT* | 0.78 | 0.7 | 0.68 | 0.75 | 1 | 0.864 |
| $NT^3$ | 1 | 0.54 | 0.98 | 0.72 | 0.78 | 0.80 |
| *NTCT* | 0.714 | 0.65 | 0.64 | 0.6 | 0.801 | 0.7 |

In this situation, some modification of the regression model may reduce the impact of multicollinearity. There are many techniques that can be used to handle multicollinearity. Converting the model to the second order is useful and overcomes the multicollinearity by combining the variables into a composite variable in the model. The generator power data do not suffer from any multicollinearity; therefore, the rest of the variables should be reduced to one variable which can be represented as ($Z$). One of the most beneficial statistical techniques that can be used to shorten the three variables in the proposed model to overcome the multicollinearity is defined as follows:

$$Z = (CT_{NEW} + NT_{NEW})/OT_{NEW} \quad [9]$$

By plotting the relationships of the observed generator temperatures with the combined independent variables, and generator power values, it can be confirmed that the quadratic model is very suitable and provides better results. The final version of the proposed model is as follows:

$$GT = \beta_0 + \beta_1.GP_{NEW} + \beta_2.Z + \beta_3.GP_{NEW}{}^2 + \beta_4.Z^2 + \beta_5.GP_{NEW}.Z$$

The final results of the proposed model are very reasonable and the problem of multicollinearity disappeared. From Table 3 we see that the variance inflation factors (VIF) for all variables are very low ($< 5$), and the tolerance values should be $> 0.2$. Moreover, we find that most of the variables are significant and contribute effectively in the model since $p < 0.05$ for each variable where p is the significant measure for any variable. When $p$ is less than 5 percent, the particular variable is very significant and participates strongly in the proposed model [9].

Table III

VIF, TOLERANCE, AND THE SIGNIFICANT VALUES

| Model | Tolerance | VIF | p |
|---|---|---|---|
| GP | 0.391 | 2.559 | 0.000 |
| $Z$ | 0.249 | 2.559 | 0.000 |
| $GP^2$ | 0.903 | 1.107 | 0.000 |
| $Z^2$ | 0.263 | 3.803 | 0.000 |
| $GP.Z$ | 0.398 | 2.510 | 0.000 |

## VI. THE RESULTS ANALYSIS

To measure the adequacy and normality of the proposed model, the model residual should be analyzed. Figure 5 shows that the error term ϵ is almost normally distributed and it is very close to normal probability plot since the majority of the residual points (the difference between the predicated generator temperature values and measured generator temperature values) are approximately distributed along a straight line. Only a few points fall outside the fitted line, which does not affect the general trend of the model. The residual is plotted against the fitted values $\widehat{GT}_i$. Figure 6 confirms that the residuals are contained in a horizontal band which indicates that there are not obvious defects. The residuals versus the independent observed variable GT should also give indication that the values are perfectly normally distributed. Figure 7 and 8 confirm this analysis. The final proposed regression equation is:

$$GT = 113 + 0.0102GP + 0.00000334Z + 0.000003GP^2 - 0.000009Z^2 - 0.000003GP.Z$$

The analysis of variance emphasizes that the model does not suffer from lack of fit since the statistic value $F_{LOF} = 0.98 < F_{\alpha,df_{LOF},df_{PE}} = 1$. The model also passed the PRESS statistic test where PRESS statistic = 39.9442 < Residual sum of square = 40.

To determine which independent variable is most important, using the length scaling method for this purpose is very powerful. The standardized coefficients of the model let us obtain the fitted equation as follows:

$$\dot{GT}^0 = 0.97\,GP + 0.003Z + 0.026GP^2 - 0.003Z^2 - 0.003GP.Z$$

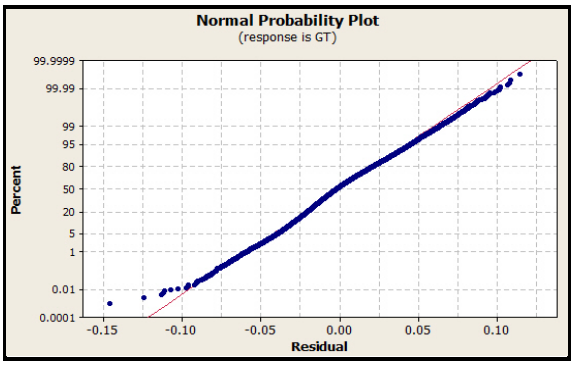


Fig.5. The normal probability plot

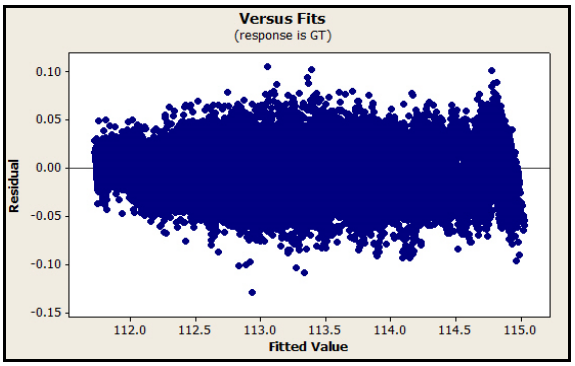


Fig. 6. The fitted values plot versus residual

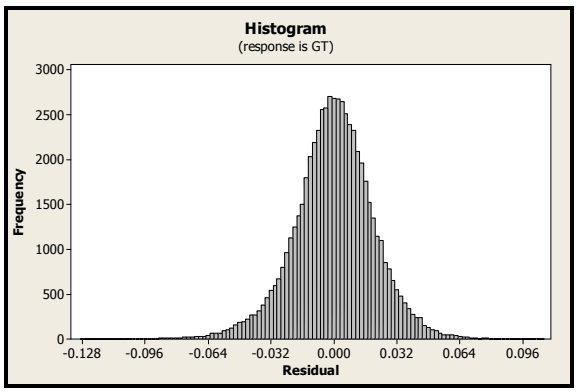


Fig. 7. The residual histogram plot

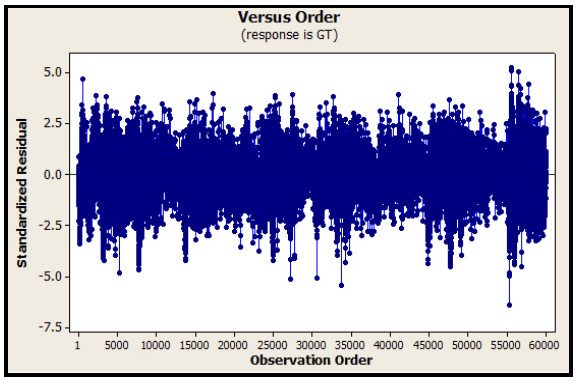

Fig. 8.The independent variable versus standardized residual

The remaining obtained results are as follows.

| | Descriptive Statistics | | |
|---|---|---|---|
| **Model** | **Mean** | **Std. Deviation** | **N** |
| ***GTT*** | 113.3583 | 1.01358 | 60000 |
| ***GP*** | .0000 | 98.56065 | 60000 |
| ***Z*** | -403.3393 | 1413.25587 | 60000 |
| $GP^2$ | 9714.0401 | 8515.30645 | 60000 |
| $Z^2$ | 2159941.4696 | 7270987.03276 | 60000 |
| ***GPZ*** | 16623.6089 | 190923.15862 | 60000 |

| | Coefficients | | |
|---|---|---|---|
| **Model** | **95% Confidence Interval Lower Value** | **Upper Value** | **Stand. Coff.** |
| ***GP*** | 0.01 | 0.020 | 0.996 |
| ***Z*** | 0.00001 | 0.000036 | 0.003 |
| $GP^2$ | 0.000002 | 0.0000035 | 0.026 |
| $Z^2$ | 0.000008 | 0.0000095 | -0.003 |
| ***GP.Z*** | 0.000002 | 0.0000035 | -0.003 |

We conclude that increasing the standardized value of generator power by one unit increases the standardized value of generator temperature by 0.97 units (dimensionless regression coefficient). Furthermore, increasing the standardized value of *Z* variable by one unit increases the standardized value of generator temperature by 0.003 units. Therefore it indicates that the generator power is more important than the generator cooling, nacelle temperatures, and outside temperatures.

## VII. CONCLUSION

Based on investigations in this paper, Multiple Linear Regression analysis can be used for analyzing multifactor data and modeling the relationship between variables. The selected variables have a direct and strong effect with the generator temperature which leads to achieve reliable results. Measuring the correlation between the observed values and the predicted values of the criterion variable based on the historical generator temperatures is the main idea of this technique. The results confirm that the polynomial regression technique is the best model that fits the obtained data. The obtained results emphasize that the generator power is the most effective variable on the generator temperatures based on the standardized coefficients of the model. The method has the advantage of being simple computationally and conceptually. Therefore, application in condition monitoring of the wind turbine generator condition can be determined by using the proposed method. With the effective selection of the data at the normal and emergency operation, this method can achieve reasonable results to predict wind turbine generator's temperature under different loads. Future work is also required to apply this method to other operational wind turbines for detection of a faulty condition and prediction of potential failure some time in advance.

## ACKNOWLEDGMENT

This work was supported in part by NSF Grant 0844707 and in part by the U.S. Department of Energy under Contract No. DE-AC36-08-GO28308 with NREL.

The authors gratefully acknowledge the help of Dr. Kathryn Johnson, who collects the data measurements from Colorado School of Mines and National Renewable Energy Laboratory.

## REFERENCES

[1] Avelino J. Gonzalez, M. Stanley Balowin, J. Stein, and N. E. Nilsson, "Monitoring and Diagnosis of Turbine-Driven Generator," *Electric Power Research Institute*, Prentice Hall, Englewood Cliffs, New Jersey 07632,1995.

[2] Olimpo Anaya-Lara, Nick Jenkins, Janaka Ekanayake, Phill Cartwright, and Mika Hughes, "Wind Energy Generation Modeling and Control," *Library of Congress Cataloguing-in-Publication Data,* 1st Edition. United Kingdom, Wiley, 2009.

[3] Bredan Fox, Damian Flynnm Leslie Bryans, Nick Jenkins, David Milborrowm, Mark O'Malley, Richard Watsn, and Olimp Anaya-Lara, "Wind Power Integration," *IET Power and Energy Series 50,* The Institution of Engineering and Technology, 1st Edition, London, United Kingdom 2007.

[4] T.W. Verbruggen, (2003). "Wind Turbine Operation & Maintenance based on Condition Monitoring," WT-Ω. Final Report, *ECN Wind Energy*, April, 2003.l. http://www.ecn.nl/ docs/library/ report/2003/ c03047.pdf.

[5] P. Guo, D. Infield, and X. Yang "Wind Turbine Generator Condition-Monitoring Using Temperature Trend Analysis," *IEEE Transactions on sustainable energy*, VOL. 3, NO. 1, JAN. 2012.

[6] W. Yang, P.J. Tavner, and M.R. Wilkinson, "Condition Monitoring and Fault Diagnosis of a Wind Turbine Synchronous Generator Drive Train," IET Renew. Power Gener., 2009, Vol. 3, No. 1, pp. 1–11.

[7] Lucian Mihet Popa, Birgitte-Bak Jensen, Ewen Ritchie, and Ion Boldea, "Condition Monitoring of Wind Generators," IEEE Industry Applications Society 38th Annual Meeting, IAS'03, Salt Lake City, Utah USA, *IEEE Signal Processing Society*, October 2003, Vol. 3, pp. 1839-1846.

[8] Data of a variable speed wind turbine 450 KW rated power, three phase permanent magnetic type 440/660 V 60 Hz. Provided by Dr. Kathryn Johnson, Colorado School of Mines, 2013.

[9] Douglas C. Montgomery, Elizabeth A. Peck, and Geoffrey Vining, "Introduction to Linear Regression Analysis," WILEY, Fifth edition 2012.

[10] Achinity Haldar, and Sankaran Mahadevan, "Probability, Reliability, and Statistical Method in Engineering Design," WILLY 2000.

[11] Minitab 16 software, http://www.minitab.com

[12] IBM SPSS software, http://www-01.ibm.com/software/ analytics/ products/statistics/ index.html

[13] J. F. Manwell, J.G.Mcgowan, and A.L.Rogers, "Wind Energy Explained Theory, Design, and Application," WILLY, Second edition 2009.

[14] Hamid A. Toliyat, Subhasis Nandi, Ngdeog Choi, and Homayoun Meshgin, "Electric Machines Modeling, Condition Monitoring and Fault Diagnosis," *Taylor & Francis Group,* 2013.

[15] Wenxian Yang, P. J. Tavner, and Michael Wilkinson, "Condition Monitoring and Fault Diagnosis of a Wind Turbine with a Synchronous Generator using Wavelet Transforms," *IET, International Comference, PEMO*, York April 2008.