Filling Missing Value Method for Power Quality Data Based on Correlation Analysis

Yundan Liang ², Wei Jiang ², Weiwei Liu¹, Shuya Lei ¹* and Weijiang Lu ¹

¹ Artificial Intelligence on Electric Power System State Grid Corporation Joint Laboratory, Global Energy Interconnection Research Institute Co. Ltd, Beijing, China.
² State Grid Corporation of China.
*Corresponding Author Email: leishuya@geiri.sgcc.com.cn

Abstract. A large number of monitoring indicators own strong correlation among them which help to better fill missing values in these sensor data. In this study, we propose an electric power quality data filling method based on correlation analysis. Firstly, normalized mutual information method is applied to deal with nonlinear correlation which makes up for the deficiency that the traditional Pearson correlation coefficient. Additionally, the measurement of correlation is calculated to obtain the closely correlated indicators. This study utilizes the regression model to build the strong regression model. Experimental results show that the approach can effectively improve the accuracy of filling, reduce the filling error, and improve the quality of data.

1. Introduction
With the fast development of the electric power Internet of things, a large number of power sensing data are gathered. However, in the process of data acquisition and data transmission, serious data quality problems existing in the large volume of monitoring data obtained. The reasons lie in the complex composition of the storage system, the harsh physical environment and the reliability of data transmission.

These "dirty" data will cause a big deviation in all kinds of data process tasks. Hence, it is necessary to clean the data and correct the wrong data before the task of data analysis. For instance, the standard electric power quality in the power system should be a perfect and symmetrical Sine wave curve. However, some factors will cause the waveform to deviate from the Sine, which will lead to a variety of electric power quality problems further. At present, the quality problems of power quality data mainly presented as data missing and data anomaly.

Moreover, data with missing value can also be named incomplete information and it is difficult to achieve accurate data analysis results based on the missing data. Due to hardware and software reasons, there are many data missing problems in power quality monitoring data. A variety of data missing situations existing and they are caused by data acquisition and network transmission problem. The existence of empty data will lead to data analysis deviation and the analysis results cannot reflect the real situation of operation. Besides, the early warning mechanism fails to work which affects the decision-making of power quality governance.

Therefore, it is imperative to cope with the serious data missing problem. With the aim to rectify data missing, we reviewed both domestic and foreign research work on the missing value filling methods. A local weighted regression-based method is proposed and are validated on real data set. The experimental results show our method could fill these missing values with a high accuracy.
2. Related Work

Many works have been proposed to solve this problem during the past decades.

The simplest way to handle missing value is to delete them. However, if there are many missing values in this dataset discontinuously, many important information are also deleted.

Another widely used method is statistic based approach such as mean value filling. In this method, mean value or the mode value of the whole time-series are taken to represent the missing value. Nevertheless, mean value filling method changed the original distribution of data by removing the random values. Its upgraded method known as weighted average method is proposed which gives each attribute a weight. Although this upgraded method owns a better performance in many situations, but how to choose a suitable weight for each attribute is another difficult problem.

What is more, there are many predict-based filling methods. Dale [1] proposed the Bayes formulation which is widely used in missing value filling methods. Especially, it works very well when all the attribute are distributed independently. Rubin [4] present imputation method based on Bayes estimation. For each missing value, this method generate m new data and the final imputation value is computed based on the processing of the m values.

Furthermore, the autoregressive analysis method [2] uses the same variable, such as the previous periods of X. An assumption is made that there is a linear relationship between these data and the missing value is estimated by its previous t-1 values.

Indeed, regression analysis method [3] is applied to train the regression model for the data without missing data. In addition, the Y attribute is predicted through the model and the value of X attribute. Regression analysis makes up for the deficiency of autoregressive analysis, but this method will also lead to the increase of the correlation between the two attributes. In the field of electric power, regression analysis method has been widely used, and the effectiveness is relatively good. In work [4], regression analysis was applied to analyze the influence of electrical environment on the power quality level of public low-voltage power grid, and the long-term trend of power quality continuous parameter time series was identified and quantified. Moreover, in work [5], in order to reduce the harmonic distortion, the quadratic regression model of multivariate variables was applied to study the harmonic interference in distribution system.

In addition, ID3 algorithm and its upgraded algorithm C4.5 [6], hot card filling [7] are also widely used. These filling methods could also achieve good results in different situations.

Recently, more works considering complex mathematic model to fill the missing values. In work [8], a smoothing cubic spline is proposed to fill the missing values of power load data. In order to solve the problem that the spline curve is not enough to represent the details of the load curve, the missing value is filled by convex combination of cubic spline and the value reflecting the mode of load curve. Reference [9] derived an interpolation method based on sinusoidal curve according to the equal interval ordered sampling value model. Although, the performance is well, but the electric power quality data of power grid is not in the order of equal interval, so it is not suitable for this scene.

3. The Cleaning Method of Power Quality Data Based on Correlation Analysis

The power quality monitoring data is obtained by the monitoring terminal monitoring power quality. The number of power quality indicators may be up to thousands where complex correlation existing among them. Given the special characteristics of the data, the power quality monitoring data can be filled based on the correlation analysis. Specifically, how to find the most relevant indicators is the key problem. In this study, a correlation-based data filling method is proposed and verified.

3.1 Indicator Selection

The metric selection problem can be viewed as a feature selection task in this study. The traditional Pearson correlation coefficient can only measure linear correlation and is greatly affected by anomalous data. We uses mutual information [10] to measure the non-linear correlation for these features. Additionally, with the aim to eliminate the impact of data volume on mutual information, the mutual information is normalized.

The original data of power quality monitoring consist of four tuples stored in JSON format. Each tuple is composed by [monitoring point, measurement index, time stamp, monitoring value]. Given the
monitoring location of each data file are the same, so the other three tuples are mainly maintained and used. Two sets of data are combined into one data set through timestamp. The correlation formula of mutual information was used to calculate and normalize the mutual information between the two indicators. The calculation process of normalized mutual information is shown in Figure 1.

\[ I(X;Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log_{\frac{p(x_i/y_j)}{p(x_i)}} \]  

**Figure 1.** The computing process of NMI.

The calculation process of normalized mutual information is as follows:

1. The data of all indicators are obtained from the original data. According to the sampling period, the receivable data amount Countrec is compared with the actual received data amount Countacu. If Countrec = Countacu, then there is no missing value in this monitoring indicator, otherwise the indicator has data missing.

2. Choose candidate indicators: The indicators with missing data are the indexes to be tested, and all the data which owns the sampling period are selected as candidate metrics. Moreover, for each monitoring data of the indicators to be selected, reorganize and sort the monitoring data according to the time stamp.

3. Compute normalized mutual information: The monitoring data of the indicators to be measured forms a dataset and the mutual information MI of the data set are computed following equation 1.
(4) Compute the max NMI: In order to eliminate the impact of the data volume, the normalized mutual information NMI is used. The indicator with the largest normalized mutual information NMI was selected as the relevant indicator.

3.2. Locally Weighted Regression [11]

The indicators with strong linear correlation can be filled in the missing values through the model established by linear regression. While for the measurement which owns weak linear correlation, other regression models were taken to fill the missing values. The power quality monitoring data has a large amount of data. In this case, if one regression model cannot fit the overall data well, the accuracy of the regression can be improved by selecting the piecewise regression method [12]. In this model, some segmentation points are selected based on the local weighted regression method.

The difference between local weighted regression and ordinary regression lies in the weight of variables. In local weighted regression model, the importance of variable is represented by the weight used.

Ordinary regression algorithm find minimum value of \( \sum_{i} (y^{(i)} - \theta^T x^{(i)})^2 \) by computing a suitable \( \theta \). The local weights are added to find minimum value of \( \sum_{i} w^{(i)} (y^{(i)} - \theta^T x^{(i)})^2 \) by computing a suitable \( \theta \).

In both of the above optimum objects, the predicted value are computed by \( \theta^T x \).

There are many weight computing methods, the most commonly used is Cubic weight function, \((n, m)\)-type weight function and Gaussian weighting function. Here we applied Gaussian weighting function and its equation is shown below.

\[
  w^{(i)} = \exp \left( -\frac{(x^{(i)}-x)^2}{2\tau^2} \right) \tag{2}
\]

We could easily get the conclusion that the point which is closer to the center point \( x \), the weight is bigger. The weight scope is \([0, 1]\). The parameter step size of \( w^{(i)} \) determines the attenuation rate of distance \( X \).

4. Experiment and Analysis

With the aim to evaluate the effectiveness of the method, the measured data sets of power quality are selected for experimental verification. It consists of monitoring data and time stamp of power grid monitoring data. Generally, the missing data are some elements in the data record. Nevertheless, in our study, we focus on the missing of the whole data record and ignore part value missing situation. For the four elements in a data record, monitoring points, monitoring indicators and time stamp, the accurate value can be obtained according to other data records. The missing of power quality monitoring data in power grid actually focuses on the missing and filling of monitoring value.

| Indicators                     | NMI | Pearson Correlation Coefficient |
|-------------------------------|-----|---------------------------------|
| 33rd harmonic voltage phase angle | 1   | 0.51                            |
| Effective value of current    | 0.85| 0.41                            |
| Harmonic current content      | 0.84| 0.26                            |
| Power factor                  | 0   | 0                               |

With the aim to calculate the correlation between indicators, the data selected in our experiment are the power quality monitoring data of high speed railway (Wuqing traction station) from 00:00 to 04:00 on March 3, 2018. In this experiment, one indicator is selected for regression experiment. The missing value existed in the 25th harmonic voltage phase angle. Its normalized mutual information NMI with
other indicators are as shown in Table 1. The maximum NMI obtained is 1, and the corresponding index is 33 times.

4.1. Indicator Selection

Experiment one: verifying the feasibility and accuracy of normalized mutual information to measure the correlation of power quality monitoring data. Five hours of data were selected for the experiment and the correlation coefficients of the results were compared.

As shown in Table 1, there is no strong correlation between 25th Harmonic Voltage Phase Angle with other indicators from view of the Pearson correlation coefficient. In contrast, from the view of NMI, we found an opposite conclusion and there is strong correlation between 33rd Harmonic Voltage Phase Angle and 25th Harmonic Voltage Phase Angle. Through the experiments conducted later, there indeed existing a strong non-linear relationship between them. Hence, the NMI could demonstrate relationship better than Pearson correlation coefficient.

At the same time, the correlation among all the whole 2555 indicators are calculated by normalized mutual information (NMI). The values of NMI > 0.8 (measurement index linear correlation, normalized mutual information value of 0.8, Pearson correlation coefficient of 0.9) were regarded as strong correlation. The calculation results are shown in Table 2.

| Metric                                      | Number |
|---------------------------------------------|--------|
| Number of indicators                        | 2555   |
| Number of indicators with strong correlation | 1538   |

As demonstrated in Table 2, there are 1538 indicators which are closely connected with each other. It means sixty percent of these indicators are strongly connected. We notice that there may exist several max NMI for one indicator. In this case, we compute each test indicator with the multiple indicators which own max NMI. The indicator with least error will be choose to predict the missing value.

4.2. Locally Weighted Regression

Experiment two: the local weighted regression experiment is a correlation experiment based on measurement indicators. According to the correlation experiment, the 25th Harmonic Voltage Phase Angle is selected as the measurement indicator, and the most relevant indicator is the 33rd Harmonic Voltage Phase Angle.

In order to show the missing value filling effect of local weighted regression, we compared it with the Mean Value filling and Polynomial fitting method. Two metrics are applied to evaluate their effectiveness and they are Root Mean Square Error (RMSE)[13] and Average Error Percentage (MAPE).

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2}
\]  \hspace{1cm} (3)

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - f(x_i)|}{y_i}
\]  \hspace{1cm} (4)

Here \(y_i\) is the ground true value and \(f(x_i)\) is the predicted value. It is worth noting that the parameter of stride (\(\tau\)) of Local Weighted Regression is vital for the performance. Table 3 shows the results of RMSE and MAPE under different setting of stride.
Table 3. Performance for Different Stride.

| stride τ | RMSE  | MAPE  |
|----------|-------|-------|
| 1        | 28.50 | 9.4%  |
| 2        | 28.05 | 9.1%  |
| 3        | 26.85 | 8.7%  |
| 4        | 26    | 8.38% |
| 5        | 26.22 | 8.45% |
| 6        | 26.20 | 8.49% |

As the result shown in Table 3, we choose stride 4 considering its best performance. Further, we randomly choose the true value and predicted value from test data set under the parameter setting. Table 4 shows the results.

Table 4. Comparison of Real Values and Predicted Values in the Test Set.

| True Value | Predict Value | Error  |
|------------|---------------|--------|
| 252.389    | 254.506       | 0.84%  |
| 277.138    | 290.907       | 4.97%  |
| 298.983    | 283.565       | -5.16% |
| 310.755    | 286.154       | -7.92% |
| 295.41     | 287.358       | -2.73% |
| 268.949    | 291.373       | 8.34%  |
| 318.89     | 286.201       | -10.25%|
| 276.235    | 286.746       | 0.31%  |
| 269.034    | 292.773       | 8.82%  |
| 226.846    | 294.322       | 29.75% |
| 317.575    | 294.006       | -7.42% |
| 296.25     | 292.333       | -1.32% |
| 296.817    | 279.346       | -5.89% |
| 261.893    | 276.845       | 5.71%  |

From Table 4, it can be seen that over 90 percent of the data own a predict error smaller than 9 percent for arbitrarily selected 14 data. The predict accuracy can meet the engineering requirements. However, we find a very high error rate in one data which is close to 30%. Through further analysis, we find an obvious fluctuation in the original data for this test data. It is very difficult to predict this kind of missing value if an anomaly existing here.

The local weighted regression method is compared with the commonly used average filling method and linear regression filling method. Figure 2 shows the test results. It can be seen from Figure 2 that the predict value of local weighted regression basically conforms to the original data. However, the result of average filling cannot reflect the characteristics of the original data given a large variation. Table 4 shows the RMSE and MAPE for these three algorithms.
Figure 2. Comparison of Test Set Filling Results.

Table 5. Filling Results Comparison.

| Filling Method       | RMSE | MAPE |
|----------------------|------|------|
| weighted regression  | 24.7 | 8%   |
| linear regression    | 26.23| 8.42%|
| average filling      | 32.33| 10.44%|

As illustrated in Table 5, local weighted regression could achieve higher accuracy in these two metrics compared with the other two methods.

However, in the previous two experiments, the length of dataset is selected as the local weighted data length. It is well-known that, the local weighted data length has a very big influence on the experiment results. Here we repeat the experiments under different data length settings. The results are demonstrated in Table 6.

Table 6. Test Results under Different K Values.

| Step | RMSE | MAPE |
|------|------|------|
| 5    | 27.90 | 9.1% |
| 10   | 27.38 | 8.8% |
| 20   | 25.77 | 8.3% |
| 30   | 26.03 | 8.5% |
| 60   | 26    | 8.4% |

An amount of experiments find that a good performance will be achieved if the number of local weighted data length is chosen as 20.

More conclusions can be draw from upper experiments. The parameter of the length of step and local weighted data show big influences on the filling effectiveness.

According to the experimental data, parameter step 4 and weighted data length 20 have a better accuracy.

In brief, if the suitable parameters are chosen, the whole predict error rate will be smaller than 10 percent which could be applied in real engineering project.

Electric power quality monitoring data has the characteristics of high-dimensionality and high sampling frequency. At the same time, there are very strong correlation between indicators which can be captured and applied to predict missing value. We proposed an electric power quality data filling method based on correlation analysis. In this approach, the Normalized Mutual Information was
introduced to measure the correlation which made up for the deficiency of Pearson correlation coefficient. Based on the NMI, the most related indicator will be chosen as basic data set which provide useful information in the further filling process. A local weighted regression method build on this indicator is applied to predict the missing value and fill them. Experiments conducted on real data set show the accuracy of filling is good and could meet the requirements of engineering project.

5. Conclusion
Electric power quality monitoring data has the characteristics of high-dimensionality and high sampling frequency. At the same time, there are very strong correlation between indicators which can be captured and applied to predict missing value. We proposed an electric power quality data filling method based on correlation analysis. In this approach, the Normalized Mutual Information was introduced to measure the correlation which made up for the deficiency of Pearson correlation coefficient. Based on the NMI, the most related indicator will be chosen as basic data set which provide useful information in the further filling process. A local weighted regression method build on this indicator is applied to predict the missing value and fill them. Experiments conducted on real data set show the accuracy of filling is good and could meet the requirements of engineering project.

6. Acknowledgment
This work is supported by Research on data governance and knowledge mining technology of power IoT based on Artificial Intelligence.

7. References
[1] A.I, Dale. On Bayes' theorem and the inverse Bernoulli theorem[J]. Historia Mathematica, 1988, 15(4):348-360
[2] Rubin, D. B. Multiple Imputation FOR Nonresponse in Surveys[M]. Hoboken, NJ: John Willey & Sons,1987
[3] Akaike H. Fitting autoregressive models FOR prediction[J]. Annals of the Institute of Statistical Mathematics,1969,21(1):243-247
[4] M Domagk, O Zyabkina, J Meyer, P Schegner. TrEND identification in power quality measurements[C]. Power EngineeringConference. IEEE, 2015:1-6
[5] SMR Rastegar, WT Jewell. A new approach FOR suppressing harmonic disturbances in distribution system based on regression analysis[J]. Electric Power Systems Research, 2001, 59(3):165-184
[6] Thakur D, Markandaiah N, Raj D S. Re optimization of ID3 and C4.5 decision tree[C]// International Conference on Computer and Communication Technology. IEEE,2010:448-450.
[7] Kim JK,Fuller W. Fractional Hot Deck Imputation[J]. Biometrika,2004, 91(3):559-578
[8] Suarez-Farinas M, Sousa R L D, Souza R C. A methodology to filter time series: application to minute-by-minute electric load series[J]. Pesquisa Operacional,2004, 24(3):355-371
[9] Jianfeng Yi, Aibo Shi, Haitao Ji, Jiawei Zhu. Correction method of sampling data missing in digital substation based on sine curve[J]. Automation and instrumentation,2014(01):116-118
[10] Viola P, Iii W M W. Alignment by Maximization of Mutual Information[J]. international journal of computer vision, 1997
[11] Cleveland W S, Devlin S J. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting[J]. Journal of the American Statistical Association, 1988, 83(403):569-610
[12] You Lv, Jizhen Liu, Wenjie Zhao, Tingting Yang. Steady-state detecting method based on piecewise curve fitting[J]. Journal of instrumentation,2012,33(01) :194-200
[13] Weijie W, Yanmin L. Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model[J]. Iop Conference, 2018, 324:012049-