Research On Two Data Cleaning Method For Multi-Dimension Wind Speed Series Data
Yongqian Liu, Hao Zhang, Jie Yan, Shuang Han, Li Li
State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, School of Renewable Energy, North China Electric Power University, Beijing 102206, China

Abstract
Measured wind speed data on the wind turbine nacelle is a significant feature which can reflect distribution of flow field on wind farms, and it is frequently used for wind power forecasting and power curve modeling. However, the actual measured wind speed data has not been fully excavated. In this paper, two wind speed data imputation methods including k-nearest neighbor imputation (KNNI) and correlation ranking imputation (CRI) are proposed to interpolate the missing and abnormal situation in multi-dimension wind speed data. In order to compare the imputation performance of these two imputation methods, Monte Carlo method is used to validate the imputation error between the imputation value and the actual value. Research shows that KNNI method is able to consider the similarity between flow fields on different time, the imputation error of KNNI is smaller than the CSI method, which is 0.944m/s. KNNI method is more suitable for the imputation of large multi-dimension measured wind speed data sets.

Objectives
Previous studies about wind turbine speed data cleaning focused on single dimension data processing. However, with rapid growth of wind turbine numbers, the amount of data needed to process is also increasing. In the same time, the complex degree of missing and abnormal situation is increase with the data increase. Imputation method such as regression imputation, specific value imputation and interpolation may not suitable for large data set which contain more dimension and more data samples, because where and how dirty data appears is unknown. So it is necessary to proposed appropriate data imputation method to process the multi-dimension wind speed series data.

In this paper, we proposed two multi-dimension wind speed imputation method based on similarity in temporal perspective and correlation in spatial perspective. Both proposed method could complex abnormal and missing situation. Final imputation result is validate by Monte Carlo method and root mean square error.

Methods

(1) Missing and abnormal situation analysis
In actual operating environment, wind speed data missing or abnormal situations are aroused by wind meter fault, data transmission failure, data storage abnormalities and artificially data deletion, etc. In the analysis, two types of data missing (shown in figure 1) and one type of data abnormality (shown in figure 2) are discussed.

(2) K-Nearest Neighbor imputation based on flow field similarity
The idea of K Nearest Neighbors Imputation (KNNI) algorithm is that finding K nearest neighbors of samples which contain some missing value through the calculation of flow field state similarity, and use the mean value of this K nearest neighbors to replace the missing value. For the advantages of KNNI, it is more flexible than regression imputation and interpolation because it ignore where and how data missing. And it is more reasonable than fixed value imputation (mean and median). It is propitious to data cleaning task for large multi-dimension wind speed series data set.

Conclusions
In this paper, two types of data missing and one type of data abnormality are analyzed. Then imputation of multi-dimension wind speed series data is implemented based on two thoughts: similarity in temporal perspective (KNNI) and correlation in spatial perspective (CRI). KNNI method based on similarity of flow field which give the reference value more reasonably, so the imputation RMSE of KNNI is lower than CRI, and imputation series of KNNI is closer to real series than imputation series of CRI, the lowest imputation RMSE of KNNI is 0.944m/s. For a long time multi-dimension wind speed data sets, KNNI is able to imputation the complex missing and abnormal situations with high imputation accuracy and stable imputation performance.

References