# Data Cleaning for Power Quality Monitoring

Zijing Yang, Junwei Cao, Yanxiang Xu
Research Institute of Information Technology
Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China

Huaying Zhang, Peng Yu, and Senjing Yao
Shenzhen Power Supply Co. Ltd.
Shenzhen 518020, China
Email: jcao@tsinghua.edu.cn

***Abstract*—Power quality issues are becoming more critical for high-tech enterprises and grid companies. Many power quality monitoring systems are deployed in recent years. Advanced analysis of monitoring data is not widely applied due to the lackness of data management. In this work, data cleaning technology is introduced to enable advanced study of power quality data, with detailed procedures and software implementation. With power quality monitoring data from Shenzhen Power Supply Company, the effectiveness of data cleaning technology applied for power quality data analysis is demonstrated. Cleaned data that avoid voidness and lackness is more feasible in actual usage, as a good basis for further advanced analysis of power quality.**

***Keywords—Power quality; Advanced analysis; Data cleaning***

## I. INTRODUCTION

As facilities like computer, information device, precision instrument, high-end manufacturing being applied to the equipments with power supply quality sensitivity, the power quality issues are catching more and more attention nowadays. To sensitive users for example semiconductor manufacturing, the voltage sag lasting for dozens of milliseconds may lead to equipment damage, stop production of line and huge economic loss. Shenzhen, as the fourth domestic city with load that surpass ten million, its electricity consumption reaches even as much as that of a province, with most consumption is from the hi-tech enterprises. Recently, sensitive customers such as microelectronic, semiconductor, biological medicine, precision manufacturing, hospitals, finance, communication industry and large-scale data center are all requiring better power supply.

At present, some domestic areas are starting the construction of power quality monitoring system in succession, with some have achieved initial success. Two construction modes are mainly employed: introduction of mature system and independent development. For power companies of Shanghai, north China and Yunnan, the PQWiew3.2 platform from the electric power research institute of America is adopted and taken for secondary development, and the network power quality monitoring platforms are well implemented. For power companies of Guangdong, Zhejiang and Jiangsu, the independent development is utilized to investigate the regional power quality monitoring network. The power quality monitoring system of Shenzhen power grid now has taken shape, with 146 substation that have installed 651 power quality monitoring terminals, while through the monitoring system, main functions like patrol, remote operation, data query and analysis, report and statistics can all be realized. However, the advanced analysis application based on data mining is still required to be further developed.

To better support the advanced analysis application, monitoring data management and cleaning for power quality are required thanks to the following reasons: irrelevant events of power quality are recorded since the developers of monitoring system are unfamiliar with the related standards of power quality. The correlating events are repeatedly recorded as the developers of monitoring system only have poor understanding to the operation principle of transmission and distribution system. Due to the system error of communication, some recorded data may be of irrationality or inconsistency and needed to be eliminated.

The main contribution of this work lies in applying the data cleaning technology to the management of power quality monitoring data. By analyzing the specific data of power quality monitoring events from Shenzhen power grid, corresponding results are obtained and the effectiveness is therefore verified

## II. DATA CLEANING

Data's validity, consistency and efficiency are the base of advanced analysis for power quality monitoring data, which can be realized with the data cleaning technology. Data cleaning refers to the process of detecting and correcting corrupt or inaccurate records from a record set, table or database. Reasons like user entry errors, corruption in transmission or storage may result in incomplete, error or irrelevant information to record set. Hence, data cleaning is in demand and employed to efficiently identify, replace, modify or delete these dirty data so as to keep the record set in consistency with the other similar record sets of the system.

The earliest research on data cleaning abroad was conducted in the US, from the error correction of social insurance number of the whole America [1]. Later, the development of American information industry and commercial had greatly promoted the data clean research which was focused on four aspects i.e. abnormal data detecting and deleting, similar duplicate record detecting and deleting, data integration and data cleaning for particular fields. However, the domestic research on data cleaning started a little late, and was mainly applied to fields such as data warehouse, decision support, data mining, comprehensive data quality management and the like, while the data cleaning for

commercial aimed at concrete application but of poor theoretical property [2-4]. So combined with practical application, how to adopt efficient models and methods to implement data cleaning for dirty data like incomplete data, error data and repeated data in order to improve data quality, is still a difficult problem than needs to be further studied [5].

## III. ADVANCED ANALYSIS FOR POWER QUALITY MONITORING DATA

Power quality monitoring data provide good support for further investigation on practical power quality problems occurring in operating grid. At present, lots of research for these data have been done, which can be divided into two categories according to research means and application purpose: one is the fundamental analysis based on data statistics and with the goal of reaching optimization, evaluation and operation management of power quality monitoring [6-8], while the other one is the advanced analysis based on data mining and adopted to extract hidden modes and rules from large scale and high dimensional power quality monitoring data to offer basis for power system planning and decision support [9-12]. The main process of advanced analysis for power quality monitoring data is shown in Fig. 1 through which it can be found out that data cleaning is the premise of advanced analysis for power quality monitoring data.

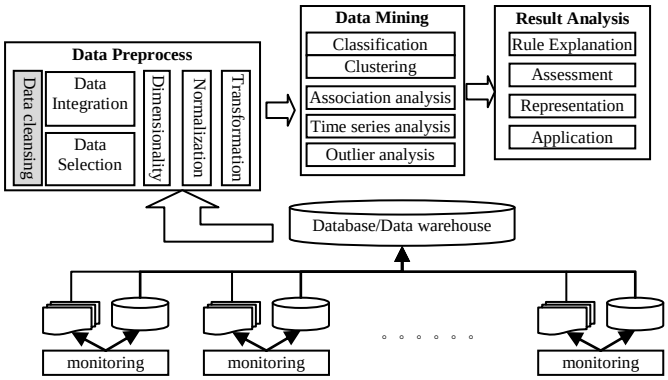


Fig. 1 Flowchart of advanced analysis for power quality

## IV. MONITORING DATA CLEANING OF POWER QUALITY

To discuss the importance of data cleaning technology applied in data analysis, the corresponding data of transient events recorded by the power quality monitoring system are especially taken to be well cleaned and managed.

### A. Concrete Steps

*a)* Data deleting. This means the data that do not match the transient events of power quality will be deleted. According to the classification standard proposed by the IEEE SCC 22 and already adopted by the IEC, voltage variation (including voltage swell, voltage sag and short time voltage interruption) with duration between half cycle and 1 minute is defined as transient event of power quality. Therefore, monitoring data that run out of this range will be deleted.

*b)* Event statistic. This refers to the times of power quality transient events will be taken for accurate statistic. Currently, the power transmission and distribution is realized by the power system consists of three-phase alternating current i.e. three ac circuits with the same frequency, same potential amplitude and phase difference of 120 ° to each other. In the statistical process, in order to guarantee the precision of statistical results, the two-phase or three-phase transient events of the same bus under the same substation and also occurring at the same time will be considered as only zero or one event according to the analysis goals (i.e. voltage swell, voltage sag and short time voltage interruption).

*c)* Data reprocessing. This means the data processed with the data deleting and event statistic steps presented above will be taken for further processing. On the basis of anticipated analysis target, some data may be of irrationality or inconsistency. So the monitoring data are requiring for further reprocessing according to relevant information such as the correlation analysis of events and user demand.

### B. Software Developing

Software for data cleaning is quite helpful for efficient preprocessing of industry data. Recently, the data cleaning software in market are commercial developed as well as developed by colleges and research institutions, while the software developing for power quality monitoring data is still at the primary stage. In this work, the analysis software of power quality transient events (as shown in Fig. 2) is well developed with data cleaning is a very important part of it.

This software is developed based on the MATLAB platform through whose powerful computing capacity and friendly man-machine interface, function modules like parameter setting, data cleaning, wavelet analysis, noise analysis and so forth of this software are well developed.

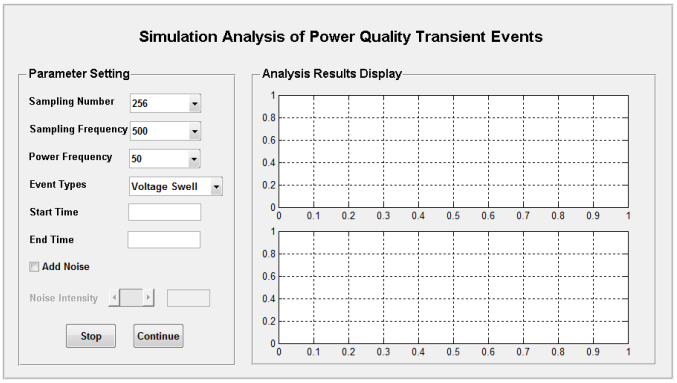


Fig. 2 Implementation of analysis software for power quality transient events

## V. EXPERIMENT VERIFICATION

In this chapter, based on the corresponding information records, the ITI (CBEMA) amplitude-time distribution maps of the power quality transient events of Shenzhen power grid is implemented. While in the maps, the CBEMA curve is the voltage tolerance curve proposed by the Computer and Business Equipment Manufacturers Association (CBEMA) according to the requirement for power quality, while the ITI

curve is put forward by the Information Technology Industry Council (ITIC) on the basis of CBEMA curve as well as of the immunity level of transient power quality required by the information industry equipments such as computers and so forth. Nowadays, the ITI (CBEMA) curve is an important basis for assessing the impact of power quality transient events and also is introduced as the American standard [13]. Plus, according to the actual situation of substations in Shenzhen, the power quality transient events are divided into substations of two levels i.e. 220kV and 110kV and then respectively employed for statistical analysis.

### *A. Case Study One*

During the operation of Shenzhen power grid, a quite outstanding problem is the voltage sag which occurs so frequently that directly leads to equipment damage, production outage of enterprises and even complaints from customers. In addition, considering there is only the amplitude distinction of rated voltage existing between voltage sag and short time voltage interruption, hence, the information records of both these two transient events of a central station in Shenzhen from 2010 to 2012 are brought for case study, while the analysis for data both before and after cleaning are executed with the following steps:

*a)* Data deleting. Firstly, as there are several central stations programmed in Shenzhen power grid, the data that are not of the selected central stations but belong to the other ones will be deleted. Then, the data with duration running out of half cycle and 1 minute will be deleted. Finally, the data of voltage swell will be deleted.

*b)* Event statistic. Since the data of three-phase circuits are all recorded, so the two-phase or three-phase transient events of the same bus under the same substation and also happening at the same time will all be considered as only one event just to avoid repeat statistic.

The analysis results are respectively shown in Fig. 3. and Fig. 4 as below:

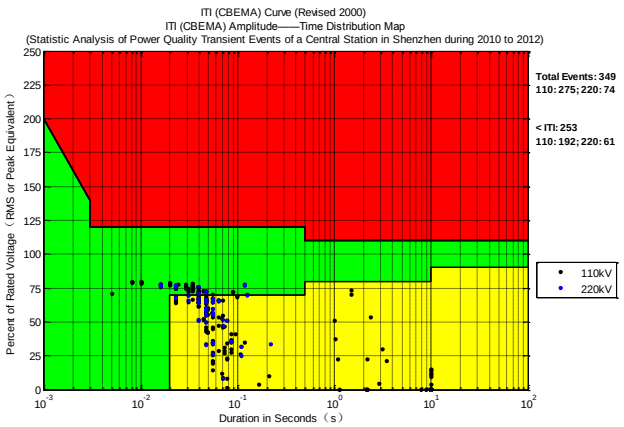


Fig. 3 Statistic results of monitoring data one of power quality before cleaning

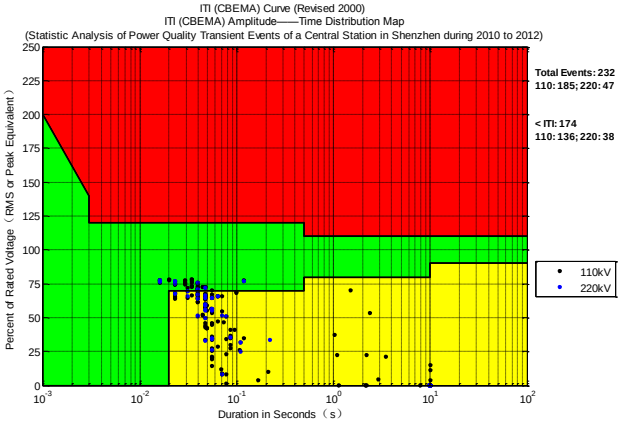


Fig. 4 Statistic results of monitoring data one of power quality after cleaning

Important data of Fig. 3 and Fig. 4 are clearly given in the table1 below:

TABLE I. CONTRAST ANALYSIS RESULTS OF MONITORING DATA ONE BEFORE AND AFTER CLEANING

| **Data Cleaning** | **Events Statistic Results** | | | | | |
|---|---|---|---|---|---|---|
| | ***Total Events*** | ***110kV*** | ***220kV*** | ***<ITI*** | ***110kV*** | ***220kV*** |
| Before | 349 | 275 | 74 | 253 | 192 | 61 |
| after | 232 | 185 | 47 | 174 | 136 | 38 |

From the two plots and table1shown above, it is not hard to find out that the total events in Fig. 4 decreases to 232 compared with 349 in Fig. 3, while the numbers of events with voltage lower than ITI decrease to 174 from 253 as shown in Fig. 3, together with the events classified into two voltage levels i.e. 110kV and 220kV present the same trend, which can be owed to the data cleaning technology.

### *B. Case Study Two*

As voltage swell is also a very typical transient power quality issue, so attention is put on this type of event herein and its corresponding information records of another central station in Shenzhen from 2010 to 2012 are taken for case study with the similar processing steps as utilized in case study one for voltage sag and short time voltage interruption, only with the procedure of deleting the data of voltage swell being replaced by deleting the data of voltage sag and short time voltage interruption, while the analysis results are respectively given in Fig. 5 and Fig. 6 as shown below:

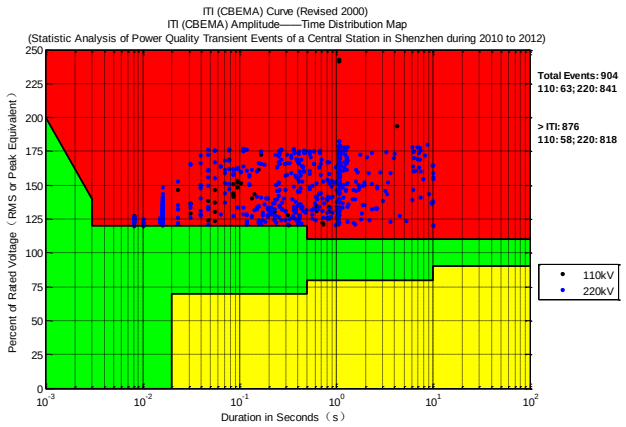


Fig. 5 Statistic results of monitoring data two of power quality before cleaning

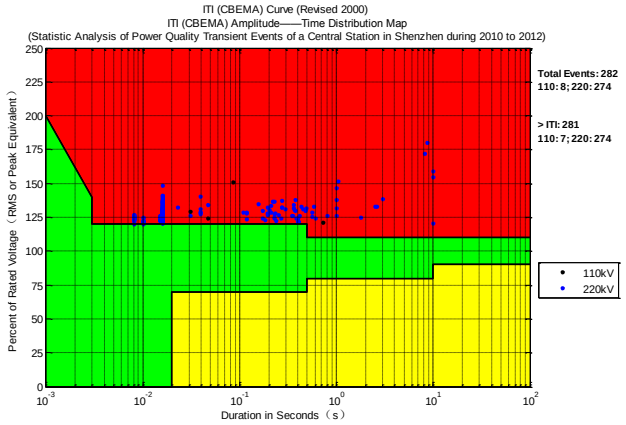


Fig. 6 Statistic results of monitoring data two of power quality after cleaning

Important data of Fig. 5 and Fig. 6 are clearly given in the table1 below:

TABLE II. CONTRAST ANALYSIS RESULTS OF MONITORING DATA TWO BEFORE AND AFTER CLEANING

| **Data Cleaning** | **Events Statistic Results** | | | | | |
|---|---|---|---|---|---|---|
| | ***Total Events*** | ***110kV*** | ***220kV*** | ***>ITI*** | ***110kV*** | ***220kV*** |
| Before | 904 | 63 | 841 | 876 | 58 | 818 |
| after | 282 | 8 | 274 | 281 | 7 | 274 |

It can be known from the two plots and table2 that compared with the total events 904 shown in Fig. 5, the total events in Fig. 6 decreases to 282, while the numbers of events with voltage higher than ITI in Fig. 6 decrease to 281 from 876 as shown in Fig. 5, together with the events classified into two voltage levels i.e. 110kV and 220kV present the same trend, which can be owed to the data cleaning technology. As for the larger difference between the results obtained respectively before and after data cleaning compared with that appears in case study one, the distinct classification of event types based on three phase circuits can be offered as the main reason.

## *C. Case Study Three*

In this chapter, information records containing all the three typical power quality transient events of another central station in Shenzhen from 2010 to 2012 are adopted for case study, and the concrete steps of data cleaning in this example are given in the following:

*a)* Data deleting. Firstly, the data that are not of the chosen central stations but belong to the other ones will be deleted. Then, the data with duration running out of half cycle and 1 minute will be deleted.

*b)* Event classification. The events records processed by step a) will be identified and classified into two categories: one is the voltage swell and the other is the voltage sag as well as short time voltage interruption in order for further separate statistical analysis.

*c)* Event statistic. The two-phase or three-phase transient events of the same bus under the same substation and also happening at the same time will be considered as only one event.

Analysis results of this case study are shown in Fig. 7. and Fig. 8 as follows:

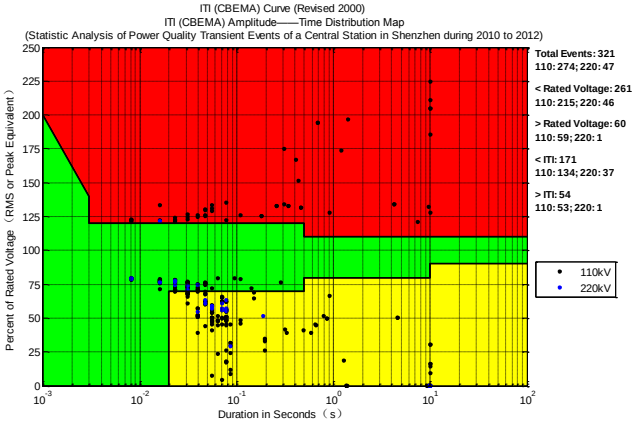


Fig. 7 Statistic results of monitoring data three of power quality before cleaning

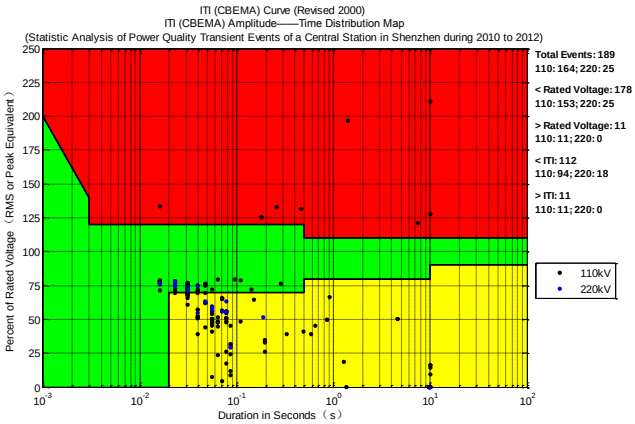


Fig. 8 Statistic results of monitoring data three of power quality after cleaning

Important data of Fig. 7 and Fig. 8 are clearly given in the table1 below:

TABLE III. CONTRAST ANALYSIS RESULTS OF MONITORING DATA THREE BEFORE AND AFTER CLEANING

| **Data Cleaning** | **Events Statistic Results** | | | | |
|---|---|---|---|---|---|
| | ***Total Events*** | ***<Rated Voltage*** | ***>Rated Voltage*** | ***<ITI*** | ***>ITI*** |
| Before | 321 | 261 | 60 | 171 | 54 |
| after | 189 | 178 | 11 | 112 | 11 |

From the results of case study three it can be seen that all the five statistic index i.e. total events, total number of voltage sag and short time voltage interruption events, total number of voltage swell events, events with voltage lower than ITI and events with voltage higher than ITI in Fig. 7 decrease apparently compared with the results got in Fig. 8, together with the events classified into two voltage levels i.e. 110kV and 220kV under each of the statistic index present the same trend, which can also be owed to the data cleaning technology.

Through the above three case studies, it can be inferred that with different anticipated analysis target being selected, different data cleaning procedures will be accordingly employed. However, one thing is still in common, that is only with the applying of data cleaning technology can more efficient and accurate results be provided.

## VI. CONCLUSION

In this work, the data cleaning technology is applied to the power quality monitoring data to better support further advanced analysis application. Due to various reasons, the big data generated by power quality monitoring system may be of kinds of problems like incorrectness and inconsistency, which may even cause misleading. If these data cannot get efficient management and deletion, the serious situation of practical power quality cannot be accurately reflected.

Combined with the actual situation of Shenzhen power grid, this work gathered important data of power quality monitoring system for further investigation to clearly illustrate the necessity of data cleaning, while the comparison of practical case respectively before and after data cleaning was also given. Shenzhen as a city that has lots of high-tech industries asks for quite high power quality. Therefore, this work provided great significance for further analysis and control of power quality issues in Shenzhen.

## ACKNOWLEDGMENT

This work is supported by the National 973 Basic Research Program of China (2013CB228206), National Natural Science Foundation of China (61233016), and China Southern Power Grid R&D project (K-SZ2012-026).

## REFERENCES

[1] H. Galhardas and D. Florescu. An extensible framework for data cleaning. In: Proceedings of the 16thIEEE International Conference on Data Engineering. San Diego, California, 2000, pp. 312.

[2] Y.F. Wang, C.Z. Zhang, B.B. Zhang and T.T. Wu. A survey of data cleaning. Information Analysis and Research. 2007,Vol. 12, pp. 50-56.

[3] J.J. Cao, X.C. Diao, T. Wang and X.F. Wang. Research on domain-independent data cleaning: A survey. Computer Science. 2010, Vol. 37(5), pp. 26-29.

[4] O. Ye, J. Zhang and J.H. Li. Survey of Chinese data cleaning. Computer Engineering and Application. Computer Engineering and Applications. 2012, Vol. 48(14), pp. 121-129.

[5] Y.Q. Jiang, Y.H. Yang and H. Yang. Investigation on impact of data cleaning to information retrieval quality and the cleaning method. Journal of the China Society of Indexers. 2012, Vol. 1, pp. 16-20.

[6] X.N. Xiao, M.X. Han and Y.H. Xu, Analysis and control of power quality. Beijing: China Electric Power Press. 2004.

[7] H.Z. Tang and J.C. Peng. Research on synthetic and qualificated appraisal index of power quality based on fuzzy theory. Power System Technology. 2003, Vol. 27, pp. 85-88.

[8] H. Jiang, J.C. Peng, Y.P. Ou and Z.Y. Li. Power quality unitary quantification and evaluation based on probability and vector algebra. Journal of Hunan University. 2003, Vol. 30, pp. 66-70.

[9] Y.S. Ou, Z.X. Song, J.H. Wang and D.G. Chen, A power quality signals de-noising algorithm based on signals' multi-scales correlation and the wavelet transform theory. Transactions of China Electrotechnical Society. 2003, Vol. 18, pp. 111-116.

[10] W.Q. Huang and Y.X. Dai. Block-Thresholding Approach for power quality disturbance denoising. Transactions of China Electrotechnical Society. 2007, Vol. 22, pp. 160-166.

[11] Z.G. Liu, Y.D. Zeng and Q.Q. Qian. Denoising of electric power system signals based on different multiwavelets. Proceedings of the CSEE. 2004, Vol. 24, pp. 30-34.

[12] L.R. Tang and X. Yang. A de-noising method of power quality based on triangle module operator. Transactions of China Electrotechnical Society. 2007, Vol. 22, pp. 154-158.

[13] IEEE Application Guide for IEEE Std 1547™, IEEE Standard for Interconnecting Distributed Resources with Electric Power Systems. 2009.