Data Quality Analysis Framework and Evaluation Methods for Power System Operation with High Proportion of Renewable Energy Penetration

Jiye Wang, Yang Li
Big Data Center of State Grid Co. Ltd
Beijing, China

Jian Guo, Junwei Cao, Haochen Hua
Research Institute of Information Technology
Beijing National Research Center for Information Science and Technology
Tsinghua University
Beijing, China
guojian_715@126.com

Zhixian Pi
State Grid Information & Telecommunication Group Co., Ltd.
Beijing, China

Caijuan Qi
Institute of Economy and Technology, Ningxia Electric Power Company
Ningxia, China

***Abstract*— Global climate crisis in 21st century pushed countries to move towards energy transformation in generation and consumption. To achieve green and low-carbon energy transformation goals, it is necessary that a large number of renewable energy resources such as wind and solar to be consumed. Renewable energy with intermittent fluctuations in time dimension and agglomerations in spatial dimension increases the complexity of green energy consumption friendly. Therefore, comprehensive data and advanced predictive analysis methods are required to guarantee safety of operation and transactions for renewable energy plants and stations. We can even say that quality of renewable energy data determines the accuracy of prediction and analysis. Firstly, this article analyzes the operation and transaction characteristics of distributed renewable energy plants, and data quality analysis framework for distributed renewable energy operations and transactions was built on the new energy cloud platform. Data information were classified into model parameter and status instance, which are related to dispatching and energy power transaction businesses such as equipment model management, operation monitoring and security analysis, measurement statistics etc. The importance between them is determined according to pairwise comparison. Finally, analytic hierarchy process (AHP) theory was applied to calculate weights for data integrity, accuracy, consistency and timeliness, data quality assessment process and calculation methods were designed, and load series data was used to verify its correctness.**

Keywords—Renewable Energy Cloud, Data Quality Analysis, AHP, Evaluation Metrics

I. Introduction

Since the beginning of the new century, the rapid development of new renewable energy sources such as wind, light, and water are gradually reducing human energy dependence on fossil. Foresighted countries such as European Union, United States, and China have successively proposed radical targets to achieve 100%, 80%, and 50% to 70% of renewable energy sources in the energy supply structure by 2050[1]. The Internet or Internet of Things (IOT) is a network designed to connecting everything with identifiers. It relies on a variety of communication modes and adopts hierarchical design structure to shield the complexity of network protocols at the bottom[2]. It provides great convenience for people to obtain data and information. While changing the previous way of communication and information exchange, It also reshapes the production and operation modes of many traditional industries [3-5]. These technological achievements have also helped energy industry achieve its structural adjustment goals by the middle of this century. Through combining big data [6] and artificial intelligence technologies with renewable energy power generation technologies, it can achieve the goal of clean and low-carbon energy. And it also can improve the efficiency of energy utilization and fairness of energy trading, and will build future smart energy systems in an efficient way [7] [8].

Various types of data are generated in different business processes of power energy, such as production and operation, equipment operation and maintenance, power consumption and transactions, which play a fundamental supporting role for the safe of power energy operation and transaction analysis. It is necessary to carry out data quality analysis and control and improve data analysis capabilities. Scholars have carried out related research. Data problems was surveyed in detail and classified into identified dimensions, and the problems are mapped into a subset of dimensions for further standardization in [9]. Then an effective data pre-processing model is proposed for processing of the big data in [10], which using relief algorithm and fast mRMR together as a hybrid approach. A distributed parallel process modeling approach is presented based on a MapReduce framework for big data quality in [11]. A pre-processing framework was proposed to address quality of data in a weather monitoring and forecasting application in [12]. The electricity consumption data was studied carefully in [13], which focuses on the data quality and outlier detection of electricity consumption data. By establishing various basic models and adjusting the collected data as observations with the same accuracy, the data quality is evaluated and corrected through adjustment or general sampling principles, which is also a common method for data quality control.

Different from the traditional power generation modes such as thermal and hydro, renewable energy power generation with wind and solar energy are characterized as intermittent and fluctuating[14]. The weather condition affects its power output directly. So improving the capacity of new energy consumption depends more on advanced power prediction technology, which requires more accurate, comprehensive and detailed real time or historical data analysis. Combining qualitative analysis with quantitative evaluation to discover potential data quality issues is becoming an important issue. And it is useful to improve safety level of renewable energy operation and transactions fairness.

This paper contributions are summarized as follows:

- Analysis data formats and types of renewable energy operation and transaction, maps them to a set of data quality types.
- From the view of integrity, accuracy, consistency and timeliness of data, construct quantitative analysis indexes and calculate weights using AHP theory.
- Propose calculation process and methods for data quality evaluation and verify its correctness.

The remainder of this paper is organized as follows. Section II constructs data quality framework for massive distributed renewable energy operation and transaction. Section III describes data quality problems, and construct quantitative analysis indicators and calculate their weights. Section IV case study. Finally, conclusions are made.

II. Renewable Cloud BASED Data quality Analysis and Evaluation

A. Architecture Design

Data quality analysis framework for large-scale distributed renewable energy operations shows in Figure 1. According to the characteristics of regional resource endowment and construction of energy transmission channels, renewable energy such as wind farms and photovoltaic power stations were built in reasonable way. Data and files generated from system operation of stations will be uploaded to renewable energy cloud platform in real time to storage uniformly and global sharing. Renewable energy data quality analysis and control includes four parts: data source analysis, data preprocessing, data quality evaluation, and data quality analysis.

Fig. 1. Renewable energy cloud based data quality analysis for massive renewable energy

1) Data Source Analysis

Its function is to classify and identify data sources generated by renewable energy plants, including domain analysis and data type analysis. Domain analysis divides data into three types, namely internal data, external data and model data. Internal data comes from production and operation of the renewable energy power plant, which provides decision support for dispatching agency and accepts control management from superior dispatching agency. External data is closely related to production and operation of renewable energy stations. Meteorological environment data provides high-precision weather data for power scheduling plans. Model data is divided into two levels, namely station model and equipment model. Station model provides accurate parameters and topology for regional-level online analysis and evaluation applications. Equipment model provides parameters and topology for station-level monitoring and security analysis. Type analysis divides data format into structured, unstructured, and semi-structured. Different reading modes are used to obtain corresponding data type and form data quality analysis instance into database.

2) Data Preprocessing

Periodic and eventual data, as well as equipment model data are maintained and corrected in this part. Its process includes data cleaning, data integration, normalization and storage management. Data preprocessing provides accurate data for upper application analysis of renewable energy operations and transactions timely.

3) Data Quality Evaluation

Evaluation indicators for renewable energy data are established based on four dimensions about data integrity, accuracy, consistency and timeliness. It has two processes that pre and post-evaluation. Pre-evaluation is that evaluates data quality before revision, and post-evaluation is that evaluates data after revision.

4) Data Quality Analysis

There are two kinds of analysis type that classification and cluster analysis. Data mining methods such as association analysis, time series analysis, and outlier analysis are used to discover abnormal or unreasonable data, and meanwhile provide calculation basis for subsequent data repair.

B. Data types in Renewable Energy Operation and Transaction

For renewable energy operation and transaction, there are mainly two types of data: static model data and dynamic instance data.

1) Model parameter data

It refers to physical information of renewable energy stations, and mainly includes three types of data, namely equipment parameters, geographical location, and transaction configuration. Equipment parameters are data of model in equipment, such as photovoltaic modules, wind turbine components, inverter, battery, etc. Geographical location information contains site name, location, ownership, capacity, link relation, affiliated organization, voltage level, accommodation way, unified/non-uniformly flag. Transaction configuration contains mid-to-long term, day-ahead and real-time generation schedule plan, transaction mode and type, transaction evaluation.

2) Status instance data

It refers to data generated during operation and transaction of renewable energy stations, mainly characterized in time dimension. External environmental sequence data contains wind speed, illumination intensity, ambient temperature, humidity, sunshine duration, gear speed, component temperature, oil temperature, oil pressure, output power, output voltage, output current. Control status data contains action instructions, alarm codes, switching/cutting, and fault recording information under emergency and fault conditions, which are all related to power energy production. Market transaction data contains spot prices, transaction power, transaction price and time, credit evaluation, and also have renewable energy-related equipment price, stock price, futures, options of manufacturer.

III. METHOD AND METRICS

A. Evaluation Method

Data quality evaluation process consists of seven parts, namely load data, integrity analysis, consistence analysis, timeliness analysis, accuracy analysis, weight calculation and store data. The diagram shows in Figure 2.

Fig. 2. Renewabl energy data quality analysis diagram

Data integrity analysis is to count missing records. The formula for calculating completeness rate is:

$$Q_1 = 1 - \frac{S_1}{S_2} \qquad （1）$$

Where $S_1 = \sum Lines\ of\ LossData$， $S_2 = \sum Lines\ of\ grossData$

Consistency analysis is matching analysis that compares attribute definition of table with data record format. The formula for calculating consistency rate is:

$$Q_2 = \frac{C_1}{C_2} \qquad （2）$$

Where $C_1 = \sum column\ of\ Data\ file,\ C_2 = \sum column\ of\ attributes$

Timeliness analysis is statistical records of data records update. The formula of update rate is:

$$Q_3 = \frac{U}{S_2} \qquad （3）$$

Where $U = \sum lines\ of\ updateData$

Accuracy analysis is to reflect the true extent of renewable energy data records. The calculation process is:

(1) Obtain recent historical data(one month or a year);

(2) Cluster analysis that using K-means algorithm[17] to calculate feature curve of series data;

(3) Calculation accuracy rate;

$$Q_4 = 1 - \frac{O}{N} \qquad （4）$$

Where $O = \sum data\ out\ of\ characteristic\ curve$ ，$N = \sum number\ of\ lineCurveData$.

Then calculating data quality is:

$$\mathrm{Q} = w1 * Q_1 + w2 * Q_2 + w3 * Q_3 + w4 * Q_4 \qquad （5）$$

Where *w1* stands for integrity, *w2* stands for accuracy, *w3* stands for consistence, *w4* stands for timeliness.

B. *Evaluation Metrics*

AHP theory was used to construct evaluation system[17]. It consists of three level, first level is mode, second level is instance , the third is metrics level. Overall hierarchy shows in Figure3.

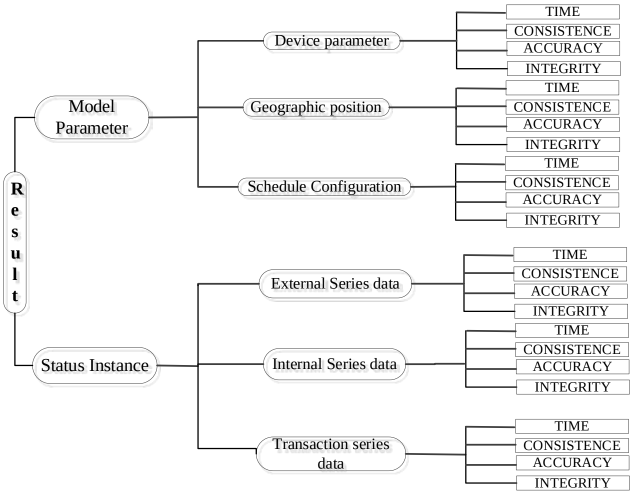

Fig. 3. AHP based evaluation hierachy

Model parameter and status instance data, are two modes of data quality evaluation. They focus on static and dynamic operating scenarios of renewable energy systems respectively. Data coupling relationship between them is relatively independent. Weighted average method was used to obtain result. Model parameter mode consists of three types of equipment parameters, geographic location, and schedule configuration. They have different requirements for data integrity, accuracy, consistency, and timeliness. State instance mode includes external environment data, internal production and operation data, and market transactions data. Classes have the same data quality requirements.

The importance between them are quantified in Table I. Importance of integrity emphasizes whether data is adequately recorded and has not been lost. Importance of accuracy emphasizes whether data is correctly recorded and reflects true value. Importance of consistency emphasizes whether data conforms to standard format and consistence with definition of attributes. Importance of timeliness emphasizes whether data records are update in time and reflect data changes.

TABLE I. IMPORTANCE ANALYSIS OF RNEWABLE ENERGY DATA EVALUATION

Mode Classification	**Types**	**Integerity**	**Accuracy**	**Consistence**	**Timeliness**
Model Parameter	Device parameter	Strongly important	Extremely	More	Slightly
	Geo position	Strongly	More	General	General
	Schedule configuration	Strongly	Extremely	More	Slightly
Statuts Instance	External - enviornmental serial data	More	More	Slightly	Strongly
	Internal-operating serial data	Strongly	More	Strongly	More
	Transaction serial data	More	Strongly	Strongly	Strongly

Under renewable energy data cloud, mode of model parameter is static description of renewable energy stations. It is foundation for renewable energy operation and transaction analysis, and requires higher data integrity and accuracy. Its requirements for consistency and timeliness are model's splicing and interaction standards. Mode of status instance is dynamic description of renewable energy stations. Its data comes from site-level monitoring systems such as CMS (Connected Wind Turbines Monitoring System) and SCADA (Supervisory Control and Data Acquisition)[15]. Its accuracy and integrity requirements are analysis of format and error analysis, which limited by communication bandwidth and upload frequency, and their requirements for timeliness and consistence are more than mode of model parameter.

Therefore, judgement matrices of model parameter and status instance J_1 and J_2 are as follows:

$$J_1 = \begin{bmatrix} 1 & 1/2 & 6 & 8 \\ 2 & 1 & 3 & 2 \\ 1/6 & 1/3 & 1 & 1 \\ 1/8 & 1/2 & 1 & 1 \end{bmatrix}$$

$$J_2 = \begin{bmatrix} 1 & 1/2 & 3 & 5 \\ 2 & 1 & 3 & 3 \\ 1/3 & 1/3 & 1 & 1/2 \\ 1/5 & 1/3 & 2 & 1 \end{bmatrix}$$

Calculating consistence index CI_1 = 0.08 <0.1, CI_2 = 0.09 <0.1 for J_1 and J_2, which meet consistency of test. Then, weight values of indexes are calculated and show in Table II.

TABLE II. EVALUATION WEIGHT OF RENEWABLE ENERGY DATA QUALITY INDICATORS

Classification	**Integrity(w1)**	**Accuracy(w2)**	**Consistence(w3)**	**Timeliness(w4)**
Model and Parameter	0.41	0.43	0.09	0.07
Status instance	0.34	0.42	0.1	0.13

IV. CASE ANALYSIS

Load time series data of September in Ningxia province was choosed for analysis and evaluation. Data table structure shows in Table III.

TABLE III. LOAD SERIAL DATA TABLE STRUCTURE

Table Name	Attributes	Remark
Load	month	history load data
	day	
	year	
	max	
	value	
Data	yesterday	yesterday load
	before yesterday	before yesterday load
Weather	temp	Weather information
	Wind	
	Rainfall	
	Humidity	
	Light radiance	
	Overlook	
	Date	
Plan value	load value	Load forecasting data

One-month load of historical data used to do case study. The number of records is 96 * 30, and the unit is kW. Their value curves of time shows in Figure 4.

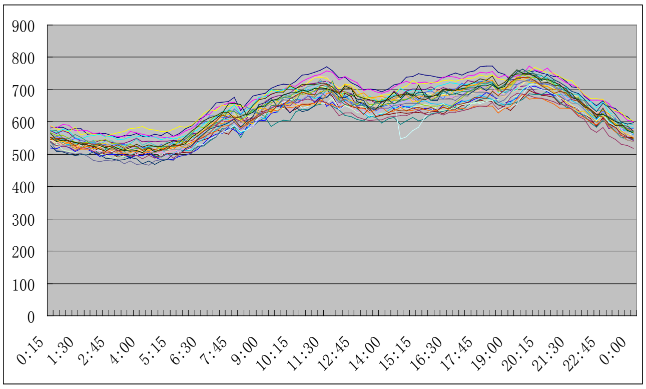

Fig. 4. Load curves of one month

Load curves are generally similar in shape except that the values are different. Larger values in the curves represent workday load, and smaller represent load of rest day. K-means algorithm was used to extract feature curve with k = 2. Two curves are calculated and show in Fig. 4.

Fig. 5. Feature curves of load data

There are two types of feature curve, one is unrepaired, and the other is repaired using interpolation method, which are identified in different colors. According to 4.2 calculation process and formulas, evaluation results of load time series data with bad data or repaired show in Table IV.

TABLE IV. LOAD DATA QUALITY EVALUATION

	integrity	accuracy	consistence	timeliness	overall
No bad data	0.92	0.98	0.95	0.98	0.95
With bad data	0.9	0.87	0.92	0.97	0.89

V. CONCLUSIONS

This paper built analysis framework of data quality for massive renewable energy operations and transactions. Under renewable energy data cloud environment, data types are categorized and decouples into static-type model parameter and dynamic-type time series data. From aspects of data integrity, accuracy, consistency, and timeliness, their impact on data quality were conducted in detail. According AHP theory, the quantitative analysis process and calculation methods were established. Load series data used to verify its correctness. Further research is using evaluation results to repair data in timely and accurate way.

ACKNOWLEDGMENT

This work was funded in part by Big Data Center of State Grid Corporation of China (Research on Quality Assessment and Governance Technology for Large-scale Distributed Grid-connected Data of Renewable Energy), National Key Research and Development Program of China (Grant No. 2017YFE0132100) and the BNRist Program (Grant No. BNR2019TD01009).

REFERENCES

[1] BP Statistical Review of World Energy[Online]. https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html.

[2] W. Richard Stevens, Kevin R. Fall, TCP/IP Illustrated Volume 1: The Protocols (2nd edition). Addison-Wesley Professional Press, 2011.

[3] Ikbal T, Rachida D, Mohamed A S. “Big Data Pre-processing: A Quality Framework”, IEEE International Congress on Big Data, pp.191-198, 2015.

[4] D. M. Strong, Y. W. Lee, and R. Y. Wang, “Data quality in context,”Communications of the ACM, vol. 40, no. 5, pp. 103–110, 1997.

[5] Kambatla, Karthik , et al. "Trends in big data analytics." Journal of Parallel and Distributed Computing,vol.74, no.72,pp.2561-2573,2014.

[6] H. Hua, Y. Qin, C. Hao, and J. Cao. “Optimal Energy Management Strategies for Energy Internet via Deep Reinforcement Learning Approach”. Applied Energy, vol.23,no.9, pp.598-609, 2019.

[7] J. Cao, K. Meng, J. Wang, et al. An Energy Internet and Energy Routers SCIENTIA SINICA Informationis, vol.44, no.6,pp. 714-727, 2014(in Chinese).

[8] Y. Ming, J. Yang, J. Cao, et al. “Distributed Energy Sharing in Energy Internet through Distributed Averaging”. Tsinghua Science and Technology, Special Section on Energy Internet, vol.23,no.3, pp.233-242, 2018.

[9] Nuno L, Seyma N S and J Bernardino. “A Survey on Data Quality: Classifying Poor Data”, IEEE 21st Pacific Rim International Symposium on Dependable Computing, pp.179-188, 2015

[10] Blessy T L, N. S Kumar. “An Enhanced Pre-Processing Model for Big Data Processing: A Quality Framework”, IEEE International

Conference on Innovations in Green Energy and Healthcare Technologies, pp.1-7, 2017

[11] Le Yao, Zhiqiang Ge. “Big data quality prediction in the process industry: A distributed parallel modeling framework”, Journal of Process Control, vol.68, no.8, pp.1-13, 2018.

[12] Ashish J, Nripendra N D. “Big data Quality Framework: Pre-Processing data in Weather Monitoring Application”, International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, pp.559-563, 2019

[13] Wen Chen, Kaile Zhou, Shanlin Yang. “Data quality of electricity consumption data in a smart grid environment”, Renewable and Sustainable Energy Reviews, vol.14, no. 75, pp.98–105, 2017

[14] M. H. Nehrir, C. Wang, K. Strunz et al. “A Review of Hybrid Renewable/Alternative Energy Systems for Electric Power Generation:Configurations, Control, and Applications”, IEEE TRANSACTIONS ON SUSTAINABLE ENERGY, vol.2.no.4,pp.392-403, 2011.

[15] Katayoun Rahbar, Chin Choy Chai, and Rui Zhang, “Energy Cooperation Optimization in Microgrids With Renewable Energy Integration”, IEEE Trans on smart grid, vol. 9, no.2,pp.1482-1493, 2018.

[16] Qin K, Pei Z. “On the topological properties of fuzzy rough sets”, Fuzzy Sets and systems, vol.7, no.151, pp.601-613, 2005

[17] Saaty T L. The analytic hierachy process[M].New York, USA:Mc Grew-Hill, 1980: 437-521.