A Method for Cleaning Power Grid Operation Data Based on Spatiotemporal Correlation Constraints

CHANGGANG WANG1,2, GANG MU2, (Member, IEEE), AND YU CAO2

1Key Laboratory of Modern Power System Simulation and Control and Renewable Energy Technology, Ministry of Education, Northeast Electric Power University, Jilin 132012, China
2School of Electrical Engineering, Northeast Electric Power University, Jilin 132012, China

Corresponding author: Changgang Wang (wangcg@neepu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant NSFC-51437003.

ABSTRACT Bad data in a power system fail to accurately reflect its state of operation. Cleaning bad data is thus necessary for the situational analysis of a data-based power system. Methods for the identification and modification of bad data based on state estimation may fail to converge in iterative calculations, and yield poor results when the errors are significant. State estimation also involves non-linear computation, which incurs a significant computational overhead and makes it difficult to efficiently handle large volumes of data. To solve these problems, this paper proposes a method for the detection and identification of bad data based on constraints on spatial–temporal correlation. In the spatial domain, we fix the equivalent power balance condition through the topology of the network. In sequential distribution, we establish data constraints based on the similarity of data among temporal sections. We then check the data in combination with the spatial–temporal correlation constraints. The proposed method was applied to an IEEE 14-bus system and a provincial grid, and the results show that it can reliably detect and modify bad data to significantly improve their quality. The proposed approach avoids non-linear calculations, is computationally efficient, and can quickly detect and modify bad data.

INDEX TERMS Power grid operation dataset, topology constraints, bad data, power error; error detection and modification methods.

I. INTRODUCTION

With the continual development of power grids, the complexity of their polytropic processes and operations has sharply increased. In 2019 alone, five countries—the UK, the US, Venezuela, Argentina, and Uruguay—experienced large-scale blackouts that had significant social and economic impacts. The currently used security system for grid operations, with the core of modeling, simulation, and fault anticipation, is thus severely challenged. The increasing uncertainty related to renewable power generation in particular poses unprecedented challenges to ensuring the secure and economic operation of power systems [1].

Following the rapid development of computing and communication technologies, increasingly sophisticated measurement systems for power systems have been developed that record massive amounts of operational data [2].

A collection of these data is generally known as the power grid operation dataset. It contains data on the active power of the grid, its reactive power, and voltage at the nodes. It provides a sufficient amount of information for analyzing the operational behavior of the power system, thus significantly enhancing the ability of analysts [3]. The power grid operation dataset can thus be used to realize a safe mode of operation of a data-driven grid [4].

We can theoretically determine the state of a power system by mining and analyzing the grid operation dataset [5], where this can help understand its behavior [6]. However, improper handling during measurement, faulty information transmission, and improper data storage can render the data erroneous, leading to bad data [7]. Detecting and correcting these bad data is a primary problem [8].

Many studies have been conducted on the detection and modification of bad data, and the relevant methods can be divided into two types: state estimation-based methods and data mining-based methods. State estimation is a fundamental
technology for advanced applications in energy management systems [9], [10], and state estimates have come a long way after years of development [11]. On the premise of obtaining the topology of the entire network and line parameters of a single time section, redundant measurements have been used to solve for the estimated data under the condition of minimizing the error by the least-squares estimation method. Bad data are then detected based on the difference between the measured and the estimated data [12]–[14]. While state estimation has contributed significantly to the safe operation of power systems, it has a few limitations: 1) State estimation is applicable only to a single time section, and does not reflect the constraints and connections among multiple sections. 2) The redundancy equations for state estimation increase the computational overhead, making it unsuitable for handling large amounts of data. 3) State estimation requires complete parameter information (network topology and line parameters). Incorrect information can lead to estimation errors. 4) Significant errors in the data may cause iterative calculations to fail to converge, which negatively impacts the identification of bad data.

Some researchers have developed time series and clustering techniques combined with data mining [15]–[17], but such work is limited to specific application scenarios and concrete goals. With the development of artificial intelligence, a large number of AI-based methods for bad data detection have been developed. Some researchers use AI methods such as deep learning (DL) and generative adversarial networks (GAN) to detect the presence of bad data in a sample. The principle is to identify bad data that is abnormal to the distribution characteristics based on the identification of data distribution characteristics. Within these scenarios, these methods have achieved high accuracy of identification of bad data [18], [19], but their effectiveness of recognition varies over different ranges, and they cannot guarantee consistent data detection over different cycles. Owing to the computational complexity of these methods, they also cannot meet the processing requirements of online temporal data.

Traditional data mining-based methods to analyze the operational behavior of systems necessitate mechanism modeling [20], and computational complexity is a bottleneck for them when rapidly processing large volumes of data. Big data-based techniques offer a different way from traditional modeling analysis to solve these problems. It involves direct and efficient access to information by using correlations in the data without relying on modeling. Power grid operational data features spatial and temporal correlations [21]. Spatially, topological relationships obtain among the various components of the power system network, the operations of each component also obeys energy conservation.

The operating status of the power system cannot be obtained directly, and we can perceive its operating state only by observing measured data. Therefore, there is a causal relationship between the operating state of the system and the operating data. If the power balance is satisfied during operation, the measured data, as a medium that reflects the power system, should also satisfy the power balance under topological constraints.

### II. RELATIONSHIP BETWEEN DATA AND SPATIO-TEMPORAL CONSTRAINTS

#### A. RELATIONSHIP BETWEEN SYSTEM STATES AND DATA

Power systems are complex artificial energy systems that operate following the law of conservation of energy. Because there are connections between the various components of the power system network, the operations of each component also obeys energy conservation.

The operating status of the power system cannot be obtained directly, and we can perceive its operating state only by observing measured data. Therefore, there is a causal relationship between the operating state of the system and the operating data. If the power balance is satisfied during operation, the measured data, as a medium that reflects the power system, should also satisfy the power balance under topological constraints.

#### B. TEMPORAL AND SPATIAL CONSTRAINT RELATIONSHIPS IN POWER SYSTEM OPERATION

The power system operates in compliance with physical constraints and the law of energy conservation at all times. It is shown in Figure 1.

Topological relationships obtain among the various components of the grid. For a normally operating grid, the node-injected power $P^\text{real}_{ii}(t_k)$ and line power $P^\text{real}_{ij}(t_k)$ satisfy the following rules:
For a node, the input and output powers of the node satisfy the power balance. The sum of the actual power associated with the node is zero.

\[ P_{\text{real}}^{\text{ii}}(t_k) + \sum_{j \neq i} P_{\text{real}}^{ij}(t_k) = 0, \quad j \neq i. \quad (1) \]

For the line, the sum of the actual power at both ends is the power loss in the branches.

\[ P_{\text{real}}^{ij}(t_k) + P_{\text{real}}^{ji}(t_k) = P_{\text{loss}}^{ij}(t_k), \quad i \neq j. \quad (2) \]

The operation of the power system is continuous in the time series. In general under normal operation, there are no significant, abrupt differences in the operational data before and after a given time section.

C. RELATIONSHIP BETWEEN BAD DATA AND CONSTRAINTS

The measured grid data \( P_{\text{mes}}^{ij}(t_k) \) may be mixed with error \( \delta(t_k) \), causing them to deviate from the actual data \( P_{\text{real}}^{ij}(t_k) \). The relationship among the measured data, actual data, and errors in the line power and node-injected power is shown in (3)–(5):

\[ P_{\text{mes}}^{ij}(t_k) = P_{\text{real}}^{ij}(t_k) + \delta_{ij}(t_k). \quad (3) \]

\[ P_{\text{mes}}^{ji}(t_k) = P_{\text{real}}^{ji}(t_k) + \delta_{ji}(t_k). \quad (4) \]

\[ P_{\text{mes}}^{ii}(t_k) = P_{\text{real}}^{ii}(t_k) + \delta_{ii}(t_k). \quad (5) \]

where \( P_{\text{mes}}^{ii}(t_k) \) and \( P_{\text{mes}}^{ij}(t_k) \) are power data for node injection and at the branch, respectively. \( P_{\text{real}}^{ij}(t_k) \) and \( P_{\text{real}}^{ij}(t_k) \) are their empirical values, and \( \delta_{ij}(t_k) \) and \( \delta_{ji}(t_k) \) are errors in the node-injection and branch powers, respectively.

The sum of node-related powers for the measured data is shown in (6).

\[ P_{\text{mes}}^{ii}(t_k) + \sum_{j \neq i} P_{\text{mes}}^{ij}(t_k) = P_{\text{real}}^{ii}(t_k) + \sum_{j \neq i} \left( P_{\text{real}}^{ij}(t_k) + \delta_{ij}(t_k) \right). \quad (6) \]

We can obtain (7) from (1) and (6).

\[ P_{\text{mes}}^{ii}(t_k) + \sum_{j \neq i} P_{\text{mes}}^{ij}(t_k) = \delta_{ii}(t_k) + \sum_{j \neq i} \delta_{ij}(t_k). \quad (7) \]

We discuss the following cases:

1) According to (7), when \( P_{\text{mes}}^{ii}(t_k) + \sum_{j \neq i} P_{\text{mes}}^{ij}(t_k) \neq 0 \) holds, this means that \( \delta_{ii}(t_k) + \sum_{j \neq i} \delta_{ij}(t_k) \neq 0 \), and \( \exists \delta_{ij}(t_k) \neq 0 \) \((j \neq i)\). It follows that bad data must exist.

2) According to (7), when \( P_{\text{mes}}^{ii}(t_k) + \sum_{j \neq i} P_{\text{mes}}^{ij}(t_k) = 0 \) holds, this means that \( \delta_{ii}(t_k) + \sum_{j \neq i} \delta_{ij}(t_k) = 0 \). The relevant measurement satisfies the physical constraint by excluding the possibility of a correlation measurement-based error cancellation. It follows that there are no bad data.

For line power, both (2) and (3) should be satisfied between the measured and the real data.

\[ P_{\text{mes}}^{ij}(t_k) + P_{\text{mes}}^{ji}(t_k) = P_{\text{real}}^{ij}(t_k) + \delta_{ij}(t_k) + P_{\text{real}}^{ji}(t_k) + \delta_{ji}(t_k), \quad i \neq j. \quad (8) \]

We can obtain (9) from (3) and (8):

\[ P_{\text{mes}}^{ij}(t_k) + P_{\text{mes}}^{ji}(t_k) = \delta_{ij}(t_k) + \delta_{ji}(t_k) + P_{\text{loss}}^{ij}(t_k), \quad i \neq j. \quad (9) \]

where \( P_{\text{mes}}^{ij}(t_k) \) and \( P_{\text{mes}}^{ij}(t_k) \) are the node-injection power and that measured at the branch, respectively, and \( \delta_{ij}(t_k) \) and \( \delta_{ji}(t_k) \) are their respective errors.

Theoretically, it must be the case that \( P_{\text{loss}}^{ij}(t_k) > 0 \). We discuss the following cases:

1) According to (9), when \( P_{\text{mes}}^{ij}(t_k) + P_{\text{mes}}^{ji}(t_k) < 0 \) holds, this means that \( \delta_{ij}(t_k) + \delta_{ji}(t_k) + P_{\text{loss}}^{ij}(t_k) < 0 \), and \( \delta_{ij}(t_k) + \delta_{ji}(t_k) < 0 \). The relevant measure does not satisfy the physical constraints, and thus \( \exists \left( \delta_{ij}(t_k), \delta_{ji}(t_k) \right) \geq 0 \). It follows that bad data must exist.

2) According to (9), when \( P_{\text{mes}}^{ij}(t_k) + P_{\text{mes}}^{ji}(t_k) \approx 0 \) holds, the physical constraint is satisfied in the vast majority of cases, ignoring a correlation measure for the eliminated case. It follows that there are no bad data.

3) According to (9), when \( P_{\text{mes}}^{ij}(t_k) + P_{\text{mes}}^{ji}(t_k) \geq 0 \) holds, \( \delta_{ij}(t_k) + \delta_{ji}(t_k) \geq 0 \), and thus \( \exists \left( \delta_{ij}(t_k), \delta_{ji}(t_k) \right) \geq 0 \). It follows that there are no bad data.

No significant difference is observed in the operating state of the power system in continuous stable operation pre and post. The real data should thus satisfy:

\[ \left| P_{\text{real}}^{ij}(t_k) - P_{\text{real}}^{ij}(t_k + 1) \right| \leq \varepsilon P. \quad (10) \]

If the above formula is satisfied, this means that \( |\delta(t_k) - \delta(t_k + 1)| \leq \varepsilon P, \) and bad data must thus exist.

We define the rate of change in power between two adjacent sections as in (11):

\[ dP(t_k) \% = \left( P(t_k) - P(t_k - 1) \right) \times 100% / P(t_k). \quad (11) \]

We calculated the rate of change in power in adjacent times series for data from the Jilin network. The results showed that the percentage of samples with a rate of change within 5% was as high as 80.6%. The fluctuation in power at adjacent times was slight. This result is valuable as a reference for identifying the quality of data at a given moment. We can set a certain threshold for the rate of change in power to establish the power data in the timing constraints as an additional condition for detecting bad data.

In summary, if the data strictly satisfy the temporal and spatial constraints, they reflect the state of the power grid. In case they do not, bad data must exist, and the available data thus do not accurately reflect the state of the power grid. Therefore, the spatiotemporal constraint can be used to detect bad data.
III. DETECTION AND MODIFICATION OF BAD DATA
BASED ON TEMPORAL AND SPATIAL
CORRELATION CONSTRAINTS

With the power balance constraint, we can detect undesirable data but cannot identify it. We need to compensate for more control samples by extending the constraint relationship to identify bad data.

A. EQUIVALENT POWER BALANCE

Any node in the grid satisfies the power balance shown in (12):

$$P_{ii}^{\text{real}}(t_k) + \sum P_{ij}^{\text{real}}(t_k) = 0, \quad j \neq i.$$  \hspace{1cm} (12)

The equation can be organized as (13):

$$P_{ii}^{\text{in}}(t_k) + P_{ii}^{\text{out}}(t_k) = 0.$$  \hspace{1cm} (13)

where $P_{ii}^{\text{in}}(t_k)$ is the sum of any combination of powers of $\{P_{ji}^{\text{real}}(t_k), P_{ji}^{\text{real}}(t_k), \ldots, P_{ji}^{\text{real}}(t_k)\}$, and $P_{ii}^{\text{out}}(t_k)$ is the sum of the complementary power combinations of $P_{ii}^{\text{in}}(t_k)$. The equivalent power model of a single node is shown in Figure 2.

When two nodes are connected, each satisfies the power balance. There may be a power difference between them. The equivalent power model is shown in Figure 3.

As shown in the figure above, where $P_{ii}^{\text{in}}(t_k) = \sum P_{ii}^{\text{real}}(t_k), m \neq j$. $P_{ii}^{\text{out}}(t_k) = P_{jj}^{\text{real}}(t_k)$, and $P_{ji}^{\text{out}}(t_k) = \sum P_{ji}^{\text{real}}(t_k), n \neq i$.

In the two-node equivalent power model, the sum of the equivalent power between nodes i and j is the loss of power of line i-j. The following formula shows this:

$$P_{ii}^{\text{in}}(t_k) + P_{ii}^{\text{out}}(t_k) = P_{ji}^{\text{real}}(t_k) + P_{jj}^{\text{real}}(t_k)$$

$$= P_{ij}^{\text{real}}(t_k), \quad i \neq j.$$  \hspace{1cm} (14)

a) When $P_{ij}^{\text{real}}(t_k) \approx 0$ holds, then $|P_{ij}^{\text{in}}(t_k)| = |P_{ij}^{\text{out}}(t_k)|$.

b) When $P_{ij}^{\text{real}}(t_k) \gg 0$ holds, then $|P_{ij}^{\text{in}}(t_k)| = |P_{ij}^{\text{out}}(t_k)|$.

When multiple nodes are connected, their equivalent equilibrium power is as shown in Figure 4.

As shown in the figure above, where $P_{ii}^{\text{in}}(t_k) = P_{ii}^{\text{real}}(t_k)$, $P_{ii}^{\text{out}}(t_k) = \sum P_{ii}^{\text{real}}(t_k), m \neq j$, $P_{ij}^{\text{in}}(t_k) = \sum P_{ij}^{\text{real}}(t_k), m \neq j$, and $P_{ij}^{\text{out}}(t_k) = \sum P_{ij}^{\text{real}}(t_k)$.

In the multi-node equivalent power model, the sum of the equivalent power between nodes i and j is the loss of power in line i-j, as shown in (15).

$$P_{ij}^{\text{out}}(t_k) + P_{ij}^{\text{in}}(t_k) = \sum P_{ij}^{\text{real}}(t_k), \quad i \neq j.$$  \hspace{1cm} (15)

The effect of power loss on equivalent equilibrium power is as follows:

a) when $\sum P_{ij}^{\text{real}}(t_k) \approx 0$ holds, then $|P_{ij}^{\text{in}}(t_k)| = |P_{ij}^{\text{out}}(t_k)|$.

b) when $\sum P_{ij}^{\text{real}}(t_k) \gg 0$ holds, then $|P_{ij}^{\text{in}}(t_k)| = |P_{ij}^{\text{out}}(t_k)|$.

In summary, when the line power loss is low, we can apply the equivalent power balance to detect bad data. According to grid operation, the percentage of line power losses in a 220 kV transmission line is 1%~3%, which is negligibly small compared with the actual transmission capacity of the line. This implies that the proposed approach is suitable for this case.

B. METHODS OF DETECTING AND CORRECTING BAD DATA

The equivalent power $P_{i,j}^{\text{in}}(t_k)$ in the measured data is composed of the true equivalent power $P_{i,j}^{\text{in}}(t_k)$ and the equivalent error $\delta_{i,j}^{\text{in}}(t_k)$. The correspondences are as follows:

$$P_{i,j}^{\text{in}}(t_k) = P_{i,j}^{\text{in}}(t_k) + \delta_{i,j}^{\text{in}}(t_k).$$  \hspace{1cm} (16)

$$P_{i,j}^{\text{out}}(t_k) = P_{i,j}^{\text{out}}(t_k) + \delta_{i,j}^{\text{out}}(t_k).$$  \hspace{1cm} (17)
\[ p_{i}^{\text{in}}(t_k) = p_{i}^{\text{out}}(t_k) + \delta_{i}^{\text{in}}(t_k). \]  
(18)
\[ p_{j}^{\text{in}}(t_k) = p_{j}^{\text{out}}(t_k) + \delta_{j}^{\text{out}}(t_k). \]  
(19)

Based on the above equilibrium constraints, we can detect bad data in the equivalent powers by comparing the equivalence errors.

We discuss the following cases:
1) When \( \delta_{i}^{\text{in}}(t_k) \gg \delta_{i}^{\text{out}}(t_k), \delta_{j}^{\text{in}}(t_k), \delta_{j}^{\text{out}}(t_k) \), then there is a large difference between the \( P_{i}^{\text{in}}(t_k) \) and \( P_{j}^{\text{out}}(t_k), P_{j}^{\text{in}}(t_k), P_{j}^{\text{out}}(t_k) \). This means that \( P_{i}^{\text{in}}(t_k) \) contains bad data, which can be modified using the following three measures:

\[ p_{i}^{\text{in}}(t_k) = \left( p_{i}^{\text{out}}(t_k) + p_{j}^{\text{in}}(t_k) + p_{j}^{\text{out}}(t_k) \right)/3. \]  
(20)

The estimated error \( \left( \delta_{i}^{\text{out}}(t_k) + \delta_{i}^{\text{in}}(t_k) + \delta_{j}^{\text{out}}(t_k) \right)/3 \) cannot be exactly zero, but the three errors used for estimation are all insignificant, thus ensuring that \( \delta_{i}^{\text{in}}(t_k) \) is smaller than \( \delta_{i}^{\text{out}}(t_k) \). Further, due to (21):

\[ \left| \delta_{i}^{\text{out}}(t_k) + \delta_{i}^{\text{in}}(t_k) + \delta_{j}^{\text{out}}(t_k) \right| \leq \left| \delta_{i}^{\text{in}}(t_k) \right| + \left| \delta_{i}^{\text{in}}(t_k) \right| + \left| \delta_{j}^{\text{out}}(t_k) \right|. \]  
(21)

The direct superposition of insignificant errors reduces the error with a certain probability. We replace the original error with the estimated error. Then, after the data have been modified, they change from the original \( P_{i}^{\text{in}}(t_k) \) to the following:

\[ p_{i}^{\text{in}}(t_k) = p_{i}^{\text{in}}(t_k) - \left( \delta_{i}^{\text{in}}(t_k) - \left( \delta_{i}^{\text{out}}(t_k) + \delta_{j}^{\text{in}}(t_k) + \delta_{j}^{\text{out}}(t_k) \right)/3 \right). \]  
(22)

2) When \( \delta_{i}^{\text{in}}(t_k), \delta_{j}^{\text{out}}(t_k) \gg \delta_{i}^{\text{out}}(t_k), \delta_{i}^{\text{in}}(t_k) \), this means that \( P_{i}^{\text{in}}(t_k) \) and \( P_{j}^{\text{out}}(t_k) \) contain bad data. If there are two quantities in the equivalent power set that are significantly out of equilibrium, it can be concluded that there are bad data in both two equivalent powers, and the remaining two measurements can be used to modify them.

\[ p_{i}^{\text{in}}(t_k) = \left( p_{i}^{\text{out}}(t_k) + p_{j}^{\text{in}}(t_k) \right)/2. \]  
(23)
\[ p_{i}^{\text{out}}(t_k) = \left( p_{i}^{\text{out}}(t_k) + p_{j}^{\text{out}}(t_k) \right)/2. \]  
(24)

Even though the estimated error \( \left( \delta_{i}^{\text{in}}(t_k) + \delta_{j}^{\text{out}}(t_k) \right)/2 \) cannot be completely zero, \( \delta_{i}^{\text{in}}(t_k) \) is guaranteed to be smaller than \( \delta_{i}^{\text{in}}(t_k) \). If the estimated error is used instead of the original error, then, after modification, the data change from the original \( P_{i}^{\text{in}}(t_k) \) to the following:

\[ p_{i}^{\text{in}}(t_k) = p_{i}^{\text{in}}(t_k) - \left( \delta_{i}^{\text{in}}(t_k) - \left( \delta_{i}^{\text{in}}(t_k) + \delta_{j}^{\text{out}}(t_k) \right)/2 \right). \]  
(25)

3) When there are three or more data items with significant differences, it is impossible to apply the above balancing constraint to identify bad data. It is then necessary to supplement the given sectional data with data from the previous iteration. Theoretically, there is little difference between pre and post data in power system operation. The data measured at both ends of the line increase and decrease at the same time. Thus, in case of a significant difference among \( P_{i}^{\text{in}}(t_k), P^{\text{out}}(t_k), P_{i}^{\text{in}}(t_k) \), \( P_{i}^{\text{in}}(t_k) \), and \( P_{i}^{\text{out}}(t_k) \) that cannot be identified by the equivalence equilibrium constraint, we need to complement the equivalence dataset from the previous section:

\[ P_{i}^{\text{in}}(t_k - 1), P^{\text{out}}(t_k - 1), P_{i}^{\text{in}}(t_k - 1), \]  
(26)

The criterion for modifying the given sectional data is to select the equivalent power data with confidence \( \min |P(t_k) - P(t_k - 1)| \) and modifying the remaining three, or to apply the data from the last instance directly instead.

The process for the identification and correction of bad data is shown in Figure 5, and the steps are as follows:

**Step 1:** We organize the line power and node power into the line-equivalent and node-equivalent power datasets, respectively, based on their affiliations.

**Flowchart of the proposed method.**

\[ p^{\text{out}}(t_k) = p^{\text{in}}(t_k) - \frac{\left( \delta_{i}^{\text{in}}(t_k) - \left( \delta_{i}^{\text{in}}(t_k) + \delta_{j}^{\text{out}}(t_k) \right)/2 \right).}{\text{ }}\]  
(26)
Line-equivalent Power Dataset:

\[ U_{Pij}(t_k) = \left\{ P_{inij}^i(t_k), P_{outij}^i(t_k), P_{jinj}^i(t_k), P_{jinj}^j(t_k) \right\} \]

Node-equivalent Power Dataset:

\[ U_{Pji}(t_k) = \left\{ P_{inji}^i(t_k), P_{outji}^i(t_k), P_{jni}^i(t_k), P_{jni}^j(t_k) \right\} \]

where \( P_{inij}^i(t_k) = P_{ij}(t_k) \), \( P_{jinj}^i(t_k) = P_{ji}(t_k) \), and \( P_{inji}^i(t_k) = P_{ji}(t_k) \).

**Step 2:** For the line-equivalent power dataset \( U_{Pij}(t_k) \), we identify bad data in \( P_{inij}^i(t_k) \) and \( P_{jinj}^j(t_k) \) \((P_{ij}(t_k)\) and \(P_{ji}(t_k)\)) according to the power balance constraint.

**Step 3:** If there are no bad data in \( P_{ij}(t_k) \) and \( P_{ji}(t_k) \), the system is observable. We can estimate the node-injected power directly from the line-equivalent power dataset, and determine if the original node-injected power contains bad data. Finally, we modify it according to Kirchhoff’s law. If bad data are present, we proceed to the next step.

**Step 4:** When there are fewer than two bad data items in elements of the line-equivalent power dataset, we can identify the double-ended measurement of the line containing bad data according to the principle of equivalence balance. After modifying the data according to the line-balancing principle, we return to Step 2 to judge the data once again. When three or more elements of the line-equivalent power dataset have bad data, we proceed to the next step.

**Step 5:** Bad data items are unidentifiable when there are three or more of them in elements of the line-equivalent power dataset. We need to complement the trusted measurements to identify the bad data. We group the trusted line measurements into a set \( L \) and proceed to the next step.

**Step 6:** We test the node-equivalent power dataset according to the power balance constraint test to determine the reliable node power, and generate the reliable power dataset \( J \). We test whether \( L \cup J \) can meet the requirements of system observability. If it can, we identify the bad data according to the line and node power sets, and return to Step 1 for another judgment of the data after modification according to the rule of system operation; if not, we proceed to the next step.

**Step 7:** If the system does not satisfy the requirements of observability, there are multiple bad data in the measured power. We then need to use measured data from the previous section to review the sectional data of the given one. By comparing \( U_{Pij}(t_k) \) with \( U_{Pij}(t_k - 1) \), we select the equivalent power data with confidence \( \min \left| P(t_k) - P(t_k - 1) \right| \), which in turn modifies the remaining three. We finally return to Step 2 for another judgment.

**IV. EXPERIMENTS AND RESULTS**

**A. COMPARISON OF DATA CLEANING EFFECTS OF DIFFERENT METHODS**

To test the feasibility of the method proposed in this paper, we used the IEEE 14-bus system as test system and state estimation as comparative method. The system contained 14 nodes and 20 lines. The required variables and parameters of the two methods are shown in Table 1. State estimation is based on the premise of knowing the line parameters of the network, which requires a variety of measurements to ensure that the system is observable so that the remaining measurements can be estimated. The method proposed in this paper, however, applies all active measurements in the network, where the number of measurements is 54. It does not require information on the network parameters, nor does it require a step to select quantities that guarantee the observable system. This significantly simplifies the pre-computation work needed.

Based on the power flow data of the IEEE 14-bus system, we simulated the daily load processing to obtain the power flow data with timing fluctuations of one minute for the recording interval as \( P(t_k) = [P_{ii}(t_k) \ P_{ij}(t_k) \ P_{ji}(t_k)]^T \). \( P_{ii}(t_k) \) is the node-injected power, \( P_{ij}(t_k) \) is the power at the head of the line, and \( P_{ji}(t_k) \) is the power at the end of the line. We mixed the error in \( P(t_k) \) to form \( P_{error}(t_k) \), and then executed state estimation and the proposed method to obtain the results of state estimation \( \hat{P}(t_k) = [\hat{P}_{ii}(t_k) \ \hat{P}_{ij}(t_k) \ \hat{P}_{ji}(t_k)]^T \) and those of topological constraint modification \( P(t_k) = [P_{ii}(t_k) \ P_{ij}(t_k) \ P_{ji}(t_k)]^T \).

We defined the overall deviations in the results of the two methods, where the overall deviation of the state estimates was:

\[
\Delta^{W}(t_k) = \sum_{i=1}^{n} \left| P_{ii}(t_k) - \hat{P}_{ii}(t_k) \right| + \sum_{j=1}^{n} \left| P_{ij}(t_k) - \hat{P}_{ij}(t_k) \right| + \sum_{j=1}^{n} \left| P_{ji}(t_k) - \hat{P}_{ji}(t_k) \right|.
\]  

That of the proposed method was:

\[
\Delta^{P}(t_k) = \sum_{i=1}^{n} \left| P_{ii}(t_k) - \tilde{P}_{ii}(t_k) \right| + \sum_{j=1}^{n} \left| P_{ij}(t_k) - \tilde{P}_{ij}(t_k) \right| + \sum_{j=1}^{n} \left| P_{ji}(t_k) - \tilde{P}_{ji}(t_k) \right|.
\]
We defined the mean deviation of state estimates as and that of the proposed method as 
\[ \mu(t_k) = \Delta(t_k) / N \], where 
\[ N \] is the number of measurements.

We used a random error for \( P(t_k) \) with a range of 1% to 10%. Then, we calculated the average deviation of the two methods for 1,000 time sections under different errors. The result is shown in Figure 6.

As shown in the figure above, the two methods had similar results in terms of error detection. However, as the mixing error increased, our method gradually outperformed the state estimation-based method.

We detected the error according to (29).
\[ \tilde{\delta}(t) = \tilde{P}(t) - P_{error}(t) \]  
(29)

The actual error was calculated according to (30).
\[ \delta(t) = P(t) - P_{error}(t) \]  
(30)

By comparing \( \tilde{\delta}(t) \) and \( \delta(t) \), we define the accuracy of identification \( \eta \) in (31).
\[ \eta = \frac{\tilde{\delta} \times 100}{\delta} \]  
(31)

The identification accuracy of the proposed method is shown in Figure 7.

As shown in the figure above, the two methods had similar results in terms of error detection. However, as the mixing error increased, our method gradually outperformed the state estimation-based method.

We apply the error data to the state estimation method, and its accuracy of detection of statistical errors is shown in Figure 9.

The result of the state estimation method were similar to those of the proposed method. The identification accuracy of the latter was lower in case of small perturbation errors but higher in case of large perturbation errors. Pairs of average accuracy of error detection between state estimation and our method for different statistical perturbation errors are shown in Figure 10.

As shown in the figure above, our method had a higher accuracy of detection than state estimation for significant errors ranging from 15 MW to 40 MW.
B. Calculating the Limitation Comparison

The computer used for the experiments was configured as follows: processor: core i7@2.60Hz; memory: 8 GB; hard disk: 128 SSD; operating system: Windows 10. We applied state estimation and our method on IEEE 14, 39, 57, and 118 systems over 1,000 sections. The calculated time consumption is shown in Figure 11.

As the system’s topology becomes larger became more complex, the computation time of our approach increased slowly, while the computation time that of state estimation increased significantly. This difference in computation time is occurred mainly since because state estimation involves a high large number of operations, whereas the kernel of our method is a simple linear calculation which that greatly simplifies the computation process and reduces the computation overhead.

C. Change in Data Quality Pre and Post Data Revision

We modified all the measured power values with significant errors based the two methods, and compared the power balance of the pre- and post-modification data of the nodes.

The power deviation at the nodes is shown in (32).

\[ \Delta P_{il}(t_k) = P_{il}(t_k) + \sum_{j \in i, i \neq j} P_{lj}(t_k). \] (32)

The mean value of the total power deviation at nodes in each section is shown in (33).

\[ W_{jd} = \frac{\sum |\Delta P_{il}(t_k)|}{N}. \] (33)

We cleaned the measured data for a provincial network using the proposed method. The average values \( W_{jd}^{se} \) of the power deviation at the node of each section before cleaning on the time series and \( W_{jd}^{se} \) after cleaning are shown in Figure 12.

The mean value of the total power deviation at the nodes was 9.22 MW while that in the modified data was 0.73 MW, a reduction by 12 times. This shows that the correlation power at the nodes was more balanced after data modification. According to the criterion for determining the bus power balance in code DL-516-2017 for the operation and administration of power dispatching automation systems, we used 10 MW as the limiting value of the power balance error at the node. The modified node imbalance measures were all less than 10 MW, indicating that they had improved data quality.

V. Conclusion

The main contribution of this paper is the proposal of a method to clean bad data concerning power grids using spatiotemporal correlation constraints to detect and modify them. In contrast with the traditional modeling approach, the proposed method applies the correlation between the data through spatial and temporal constraints. We used the IEEE 14-bus system and one year of data from a provincial network to test our method. The main conclusions are as follows:

1. Our method imposes fewer conditions than state estimation, and is not dependent on parameters of the transmission components. It does not involve non-linear operations, incurs a smaller computational overhead, and is thus more efficient.

2. After using different ratios of errors in the data, we applied the proposed method and state estimation to them. The results showed that the two methods had similar cleaning effects on bad data when the errors were small, indicating that our proposed method is feasible; when the errors were significant, the proposed method was more accurate than state estimation, which shows its superiority.

3. We applied our method to actual grid operation data, and the results showed that it can reduce power balance errors at the node. The data after data cleaning tended to be power balanced, and had better quality.

REFERENCES

[1] Y. Li, Z. Yang, G. Li, D. Zhao, and W. Tian, “Optimal scheduling of an isolated microgrid with battery storage considering load and renewable generation uncertainties,” IEEE Trans. Ind. Electron., vol. 66, no. 2, pp. 1565–1575, Feb. 2019.

[2] B. P. Bhattarai, S. Paudyal, Y. Luo, M. Mohanpurkar, K. Cheung, R. Tonkoski, R. Hovsapian, K. S. Myers, R. Zhang, P. Zhao, M. Manic, S. Zhang, and X. Zhang, “Big data analytics in smart grids: State-of-the-art, challenges, opportunities, and future directions,” IET Smart Grid, vol. 2, no. 2, pp. 141–154, Jun. 2019.

[3] C. Kang, Y. Wang, Y. Xue, G. Mu, and R. Liao, “Big data analytics in China’s electric power industry: Modern information, communication technologies, and millions of smart meters,” IEEE Power Energy Mag., vol. 16, no. 3, pp. 54–65, May 2018.
[4] S. Liu, D. Zhang, C. Zhu, W. Li, W. Lu, and M. Zhang, “A view on big data in Energy Internet,” Dianli Xitong Zidonghua/Automat. Electr. Power Syst., vol. 40, no. 8, pp. 14–21 and 56, 2016.

[5] Y. Yan, G. Sheng, R. C. Qu, and X. Jiang, “Big data modeling and analysis for power transmission equipment: A novel random matrix theoretical approach,” IEEE Access, vol. 6, pp. 7148–7156, 2018.

[6] J. Wu, K. Ota, M. Dong, J. Li, and H. Wang, “Big data analysis-based security situational awareness for smart grid,” IEEE Trans. Big Data, vol. 4, no. 3, pp. 408–417, Sep. 2018.

[7] J. Zhang, X. Zheng, Z. Wang, L. Guan, and C. Y. Chung, “Power system sensitivity identification—Inherent system properties and data quality,” IEEE Trans. Power Syst., vol. 32, no. 4, pp. 2756–2766, Jul. 2017.

[8] Y. Lin and A. Abur, “A highly efficient bad data identification approach for very large scale power systems,” IEEE Trans. Power Syst., vol. 33, no. 6, pp. 5979–5989, Nov. 2018.

[9] J. Zhao, J. Qi, Z. Huang, A. P. S. Meliopoulos, A. Gomez-Exposito, M. Netto, L. Milis, A. Abur, V. Terzija, I. Kamwa, B. Pal, and A. K. Singh, “Power system dynamic state estimation: Motivations, definitions, methodologies, and future work,” IEEE Trans. Power Syst., vol. 34, no. 4, pp. 3188–3198, Jul. 2019.

[10] M. Wang, “A multi-index comprehensive evaluation method of state estimation,” Dianli Xitong Zidonghua/Automat. Electr. Power Syst., vol. 39, no. 1, pp. 94–98, 2015.

[11] S. Dahale, H. S. Karimi, K. Lai, and B. Natarajan, “Sparsity based approaches for distribution grid state estimation—A comparative study,” IEEE Access, vol. 8, pp. 198317–198327, 2020.

[12] H. Ding, Z. Gao, D. Su, J. Zhao, N. Chang, and C. Gong, “Solution to bad data problem of phase angle reference bus for state estimation with hybrid measurement,” Dianli Xitong Zidonghua/Automat. Electr. Power Syst., no. 9, pp. 132–136, 2014.

[13] H. Zhang and W. Yi, “Distributed state estimation method for power systems based on asynchronous iteration mode,” Dianli Xitong Zidonghua/Automat. Electr. Power Syst., no. 9, pp. 125–131, 2014.

[14] C. H. Ho, H. C. Wu, S. C. Chan, and Y. Hou, “A robust statistical approach to distributed power system state estimation with bad data,” IEEE Trans. Smart Grid, vol. 11, no. 1, pp. 517–527, Jan. 2020.

[15] Z. Yang, H. Liu, T. Bi, and Q. Yang, “Bad data detection algorithm for PMU based on spectral clustering,” J. Modern Power Syst. Clean Energy, vol. 8, no. 3, pp. 473–483, 2020.

[16] J. Wang, Z. Xu, and Y. Che, “Power quality disturbance classification based on compressed sensing and deep convolution neural networks,” IEEE Access, vol. 7, pp. 78336–78346, 2019.

[17] J. Dai, H. Song, G. Sheng, and X. Jiang, “Cleaning method for status monitoring data of power equipment based on stacked denoising autoencoders,” IEEE Access, vol. 5, pp. 22863–22878, 2017.

[18] Y. H. A. C. Peng, “Research on quality assessment method for power grid statistical data based on cloud model,” Comput. Appl. Softw., vol. 31, no. 12, pp. 100–103, 2014.

[19] Y. Zhang, L. Shi, W. Lu, Z. Zhang, C. Zhao, and Q. Pei, “Realization of data management and control mechanism of line loss statistical information,” Dianli Xitong Zidonghua/Automat. Electr. Power Syst., vol. 40, no. 7, pp. 128–133, 2016.

[20] X. He, L. Chu, R. C. Qiu, Q. Ai, and Z. Ling, “A novel data-driven situation awareness approach for future Grids—Using large random matrices for big data modeling,” IEEE Access, vol. 6, pp. 13855–13865, 2018.

[21] Y. Ma, C. Huang, Y. Sun, G. Zhao, and Y. Lei, “Review of power spatiotemporal big data technologies for mobile computing in smart grid,” IEEE Access, vol. 7, pp. 174612–174628, 2019.

CHANGGANG WANG was born in Jilin, China, in 1987. He received the B.S. and M.S. degrees from Northeast Electric Power University, Jilin, China, in 2014. He is currently pursuing the Ph.D. degree in electrical engineering with Northeast Electric Power University. His research interest is power system dynamic analysis. From 2014 to 2016, he was an Assistant with the School of Electrical Engineering, Northeast Electric Power University. Since 2017, he has been a Lecturer with the School of Electrical Engineering, Northeastern Electric Power University.

GANG MU (Member, IEEE) received the Ph.D. degree in electrical engineering from Tsinghua University, China, in 1991. He is currently a Professor with Northeast Electric Power University, China. His research interests include power system stability analysis, planning and operation analysis of large-scale wind farms integrating to power systems, planning and control of energy storage, and big data/Al applications in power system operation. He is a Fellow and a Councillor of CSEE. He is also the Director of the Modern Power System Simulation and Control and Renewable Energy Technology Key Laboratory of Ministry of Education China.

YU CAO received the B.S. and Ph.D. degrees in electronic science and technology from Nankai University, China, in 2013. He is currently an Associate Professor with Northeast Electric Power University, China. His research interests include photovoltaic power generation theory and applications and network communication and integrated protection of smart substations.

* * *