Cost-Effective Bad Synchrophasor Data Detection Based on Unsupervised Time Series Data Analytics

Lipeng Zhu, Member, IEEE, and David J. Hill, Life Fellow, IEEE

Abstract—In modern smart grids deployed with various advanced sensors, e.g., phasor measurement units (PMUs), bad (anomalous) measurements are always inevitable in practice. Considering the imperative need for filtering out potential bad data, this paper develops a simple yet efficient online bad PMU data detection approach by exploring spatial-temporal correlations. With no requirement on specific prior knowledge or domain expertise, it performs model-free, label-free, and non-iterative bad measurement detection in power systems from a data-driven perspective. Specifically, spatial-temporally correlated regional measurements acquired by PMUs are first gathered as a spatial-temporal time series (TS) profile. Afterwards, TS subsequences contaminated with bad PMU data are identified by characterizing anomalous spatial-temporal nearest neighbors (STNN). To make the whole approach competent in processing online streaming PMU data, an efficient strategy for accelerating STNN discovery is tactfully designed. Numerical test results on the Nordic test system and the realistic China Southern Power Grid demonstrate the reliability, efficiency and scalability of the proposed approach in practical online monitoring.

Index Terms—Bad data detection, data analytics, spatial-temporal correlation, synchrophasor measurements, time series.

I. INTRODUCTION

With the rapid development of smart sensing and Internet of things (IoT) technologies, advanced wide-area measurement systems (WAMS) have been increasingly deployed in modern smart grids [1], [2]. By acquiring high-resolution measurement data using phasor measurement units (PMUs) in a synchronized manner, the WAMS significantly enhance power grids’ capability in situational awareness during online monitoring. In this circumstance, the PMU data quality acts as the cornerstone of many WAMS-based advanced applications, such as online dynamic stability assessment [3], wide-area event detection [4], and wide-area stability control [5].

However, due to the inevitability of sensing errors and WAMS component malfunctions, bad (anomalous) PMU data are widely witnessed in practice. For example, in China Southern Power Grid (CSG), statistical analysis by system operators shows that about 10%∼30% of PMU measurements are contaminated with bad data. Besides, as PMU measurements acquired during online monitoring come into the data centers and control rooms as data streams, potential bad PMU data should be timely filtered out to avoid undesirable data accumulation. Therefore, it is imperative to develop reliable and efficient bad PMU data detection (BPDD) schemes for practical power grids.

In terms of BPDD, the research community has made tremendous efforts to tackle this problem during power system online monitoring. Conventionally, state estimation (SE) related approaches are widely adopted to perform online BPDD. In [6], with the help of a non-linear weighted least squares state estimator, normalized residual tests are performed to identify bad data. An augmented state vector approach is proposed for SE in [7], which is able to both detect bad PMU data and improve the data quality. Recently, by classifying suspicious data into small groups and implementing largest normalized residual tests in parallel, a highly efficient SE based BPDD method is reported in [8]. Based on quadratic prediction and Kalman filtering, an algorithm is proposed in [9] to preprocess bad PMU data before performing SE. In addition, some recent PMU-based applications have taken the problem of BPDD into account to improve their applicability. For example, focusing on identifying multiple power line outages in the presence of bad PMU data, a systematic framework is developed in [10], which helps correct bad measurements simultaneously. With the dual target of enhancing system observability and bad data detection, a unified PMU placement scheme for wide-area SE is proposed in [11]. While these methods have exhibited their strength in coping with BPDD in their case studies, their reliability could be impaired in practice due to their heavy reliance on system topological information and model parameters.

In recent years, a handful of inspiring data-driven efforts have shown high potential in fulfilling the BPDD task. In [12], based on the low-rank property of spatial-temporal measurement matrices, missing PMU data are detected and recovered by solving a low-rank matrix completion problem. Similarly, the low-rank property of the Hankel structure is exploited to identify and correct bad PMU data in [13]. Recently, on the basis of the intrinsic spatial-temporal correlations between multiple PMU channels, a density based clustering method called local outlier factor (LOF) analysis is introduced in [14], [15] for BPDD. Essentially, these methods exploit power systems’ inherent spatial-temporal properties/correlations reflected in regional PMU measurements to detect potential anomalies. Compared with the afore-mentioned model-based solutions, these model-free alternatives would achieve more reliable BPDD in the presence of inaccurate topology information or parameter errors. Nevertheless, they have their own limitations. As the low-rank based approaches in [12], [13] involve complicated optimization procedures to solve the BPDD problem, their implementations are likely to be computationally expensive in practical online monitoring. While the LOF-based method [14], [15] carries out BPDD at a high speed, its online reliability depends on the preparation of a high-quality historical PMU database (with no bad data),
which requires tough class labeling efforts to screen out all the anomalous historical data.

Taking the above research gaps into account, this paper develops an efficient model-free BPDD scheme by unsupervised time series (TS) data analytics. Following the fundamental idea of exploiting spatial-temporal correlations to perform BPDD [12]–[15], a much simpler yet more efficient BPDD approach is put forward, with neither need for costly iteration nor for painstaking offline labeling. In particular, with sequential PMU measurements in a specific region integrated as a spatial-temporal TS profile, the BPDD problem is first converted to spatial-temporal anomaly detection from TS. Then, sequential BPDD is efficiently performed by identifying anomalous TS subsequences which remain far away from their spatial-temporal nearest neighbors (STNN).

The BPDD scheme proposed in this paper does not involve time-consuming offline training that is widely witnessed in most big data analytics approaches. In addition, it does not contain any complicated data processing or optimization procedure. Instead, after the completion of data acquisition, it simply and efficiently works in a “plug-and-play” fashion. Owing to this attractive feature, this scheme is capable of addressing online PMU data streams in a computationally efficient and reliable way. The main contributions and merits of this paper are outlined below.

- Based on unsupervised TS data analytics, this work develops a data-driven BPDD approach for practical power grids, with desirable model-free, label-free, and noniterative features.
- By fully exploring spatial-temporal correlations, the BPDD approach can precisely detect various types of bad data even with extremely weak anomalous behaviours.
- A fast STNN discovery strategy is tactfully introduced to perform BPDD in a highly simple and fast manner, which makes it suitable for handling real-time data streams.
- With the whole approach working in a cost-effective way, it exhibits strong applicability and scalability in practical contexts, as demonstrated in experimental tests on the real-world CSG with field PMU measurements.

The remainder of the paper is structured as follows. Section II describes the BPDD problem and formulates it as detecting spatial-temporal anomalies. Section III presents the model-free sequential BPDD approach in detail. In Sections IV and V, numerical tests are extensively carried out on the Nordic test system and the real-world CSG to test the performances of the proposed approach. Finally, Section VI concludes the paper.

II. BAD PMU DATA: SPATIAL-TEMPORAL ANOMALY

A. Problem Description

For a certain region in a given power grid, suppose $n_b$ PMUs are installed at $n_b$ neighboring buses for online monitoring. When these buses are located in a small region, they are expected to present relatively strong electrical couplings. Given an observation time window (OTW) of length $T$, PMU measurements are sequentially acquired from individual buses with a sampling interval of $\Delta t$. Sequential PMU data of a certain type of electrical quantities, e.g., voltage magnitude, are gathered as a spatial-temporal measurement matrix consisting of $n_b$ TS:

$$
X = \begin{bmatrix}
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_{n_b}
\end{bmatrix} + \begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_{n_b}
\end{bmatrix}
$$

(1)

where $x_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$ is the TS of PMU measurements obtained from bus $i$ ($1 \leq i \leq n_b$), and $n = T/\Delta t$ is the number of data points in $x_i$. In fact, $X$ can be decomposed into two independent parts:

$$
X = A + E
$$

(2)

where $A = [\alpha_1, \alpha_2, \ldots, \alpha_{n_b}]^T$ and $E = [\epsilon_1, \epsilon_2, \ldots, \epsilon_{n_b}]^T$ represent the ideally measured states of the power grid and the measurement errors, respectively.

In essence, due to the inherent networked electrical couplings between individual buses, neighboring buses generally have similar dynamic behaviors in both normal quasi-steady states and dynamic processes caused by transient events. Hence, relatively strong spatial-temporal correlations dwell in the $n_b$ TS of $A$, i.e., the sequences $\alpha_1, \alpha_2, \ldots, \alpha_{n_b}$ share significant similarities with each other. $X$ is expected to inherit such spatial-temporal correlations from $A$ provided that the measurement errors in $E$ are trivial. However, if the measurement errors become too large, the intrinsic spatial-temporal correlations cannot be sufficiently captured by the matrix $X$. Based on the definition in [14], such PMU data with large measurement errors are called bad (low-quality) PMU data. Note that the extreme case of complete PMU data loss can be also considered as bad data by filling the lost PMU measurements with zero values. In fact, as the spatial-temporal patterns characterized by bad PMU data are significantly different from the original spatial-temporal characteristics of the actual system dynamics, the corresponding sequential bad data are deemed as spatial-temporal anomalies in $X$. In this regard, the fundamental task of detecting bad PMU data can be converted to identifying spatial-temporal anomalies in $X$.

B. Illustrative Example

With the above description of spatial-temporal correlations in PMU data, how to exploit them to detect bad PMU data is further illustrated using real-world PMU data acquired from CSG. Three adjacent buses’ voltages in pre-fault, fault-on and post-fault periods are sequentially acquired by PMUs, as shown in Fig. 1. By comparing the differences of voltage profiles between the three buses, one can easily figure out there exist several bad data points in the first voltage profile. Specifically, it is observed the first bus voltage experiences data spikes at 5.94 s, 8.88 s and 12.56 s, respectively, while the other two bus voltages seem to have relatively smooth profiles at these time instants. Therefore, the three data points in the first voltage profile are detected as bad PMU data.

In fact, this simple BPDD example implicitly exploits the knowledge about spatial-temporal correlations between the
features are extremely helpful for accurately identifying bad data, i.e., they still sufficiently capture the inherent spatial-temporal dynamics present in the physical system. In contrast, bad data are likely to break the original spatial-temporal patterns of the features in the physical system. If the noise level is relatively high, how to identify bad data from noisy profiles would be very challenging. Additionally, if the event is not that severe, it may be difficult to distinguish them due to similar voltage deviations. As illustrated in [16], [17], (5) is equivalent to the normalized Euclidean distance.

![Fig. 1. Illustration of bad PMU data in a real-world power grid.](image)

### III. Sequential Bad PMU Data Detection

In the spatial-temporal measurement matrix $X$, if it does not contain bad data, the $n_b$ TS will resemble each other to some degree. Under this circumstance, each subsequence extracted from the TS would have at least one neighbor staying relatively close to itself. Otherwise, a TS subsequence including bad data would be a spatial-temporal anomaly that is significantly different from other subsequences, thus being very dissimilar to its nearest neighbor. Following these basic ideas, an efficient sequential BPDD approach by characterizing STNN is developed in this paper. Its technical details are described in the following.

#### A. Profiling STNN for BPDD

With all the $n_b$ TS concatenated one by one, the $n_b \times n$ measurement matrix $X$ is reshaped into a vectorized TS:

$$x' = [x'_1, x'_2, \ldots, x'_{k'}, \ldots, x'_N], \text{ for } N = n_b \times n$$

where $x'_k = x_{i,j}$ is the original entry in the $i$th row ($1 \leq i \leq n_b$) and $j$th column ($1 \leq j \leq n$) of $X$, with $i = \text{ceil}(k/n)$ (rounding towards $+\infty$) and $j = k - (i-1)n$. By designating a length value $m$ ($3 \leq m \leq N$), $(N - m + 1)$ subsequences of length $m$ are extracted from $x'$:

$$x'_{u,m} = \{x'_{u}, x'_{u+1}, \ldots, x'_{u+m-1}\}, \text{ for } 1 \leq u \leq N - m + 1$$

Given two subsequences $x'_{u,m}$ and $x'_{v,m}$ ($1 \leq u, v \leq N - m + 1$), their dissimilarity is measured by

$$d_{u,v} = \sqrt{2m \left(1 - \frac{M_{u,v} - m\mu_u\mu_v}{m\sigma_u\sigma_v}\right)}$$

where $M_{u,v} = \sum_{k=1}^{m} x_{u+k-1}x'_{v+k-1}$ is the dot product of $x'_{u,m}$ and $x'_{v,m}$, $\mu_u$ and $\mu_v$ are their mean values, and $\sigma_u$ and $\sigma_v$ are their standard deviations, respectively. Note that, (5) is equivalent to the normalized Euclidean distance.

For each subsequence in $x'$, by calculating its distances to all the subsequences in $x'$, a distance vector is obtained:

$$d_u = [d_{u,1}, d_{u,2}, \ldots, d_{u,N-m+1}] \quad (1 \leq u \leq N - m + 1)$$

The minimum distance values are then collected from each distance vector (with $d_{u,u}$ deleted from $d_u$ to avoid trivial minimum estimation) to form a vectorized profile:

$$p_{NN} = \left[\min(d_1), \min(d_2), \ldots, \min(d_{N-m+1})\right]$$

As $p_{NN}$ represents the collection of all the subsequences’ distances to their nearest neighbors in the spatial-temporal measurement matrix, it is thus called the STNN profile. By setting a threshold $\xi$, the subsequence $x'_{u,m}$ would be identified as an anomaly (containing bad PMU data) if its STNN distance satisfies the following condition:

$$p_{NN}(u) = \min(d_u) > \xi$$

The condition in (7) indicates that $x'_{u,m}$ is significantly different from others, with its nearest neighbor being far away from it. Hence, bad PMU data are expected to exist in $x'_{u,m}$. In order to let this criterion be adaptive to statistical characteristics of different spatial-temporal profiles, the threshold for decision is automatically determined by

$$\xi = \mu(p_{NN}) + K\sigma(p_{NN})$$

where $\mu(p_{NN})$ and $\sigma(p_{NN})$ are the mean value and standard deviation of $p_{NN}$, and $K$ is a coefficient controlling the
criterion’s sensitivity. Based on (9), ε represents a certain anomalous level, which is similar to the 3σ rule in Gaussian distributions. Empirical tests show that setting the coefficient to \( K = 6 \) generally results in highly reliable BPDD.

B. Fast STNN Discovery

The derivation of STNN in (7) involves numerous distance calculations and comparisons, especially when the total number of subsequences in \( x' \) is very large. With a brute-force manner of calculating and comparing pair-wise distances one by one, the overall computational complexity of profiling STNN would be extremely high. If not treated properly, such a heavy computational burden will deteriorate the proposed approach’s online performances in practice, where streaming PMU data need to be processed efficiently. To speed up online BPDD, a fast STNN discovery strategy is introduced in this paper based on a novel pairwise similarity search algorithm [18], [19].

The key to accelerating the procedure of STNN discovery lies in improving the efficiency of pairwise distance calculation in (5). As the mean values and standard deviations in (5) can be efficiently computed by existing commercial software such as MATLAB, the main concern is how to quickly obtain the dot product \( M_{u,v} \). In this paper, the convolution based discrete Fourier transform (DFT) and its inverse counterpart [18] are exploited to perform batch dot-product calculations for \( d_u \) in a vectorized manner. Before the DFT manipulation, two synthetic sequences are first derived from \( x' \) and \( x'_{u,m} \) by padding zeros and re-ordering elements in a mirrored way:

\[
x_p' = [x'_1, x'_2, ..., x'_N, 0, 0, ..., 0] \quad \text{N zeros}
\]

\[
y_u' = [x'_{u+m-1}, x'_{u+m-2}, ..., x'_u, 0, 0, ..., 0] \quad \text{(2N-m) zeros}
\]

Then, DFT is performed on \( x_p' \) and \( y_u' \), which yields

\[
X'_p = F(x'_p) = [X'_1, X'_2, ..., X'_{2N}]
\]

\[
Y'_u = F(y'_u) = [Y'_1, Y'_2, ..., Y'_{2N}]
\]

where \( F(*) \) denotes DFT. Based on the DFT calculation results, inverse DFT is further carried out:

\[
Q_u = F^{-1}(X'_p \odot Y'_u) = [Q_{u1}, Q_{u2}, ..., Q_{u,2N-1}]
\]

where \( F^{-1}(*) \) represents inverse DFT, and \( \odot \) denotes the dot product of \( X'_p \) and \( Y'_u \). In fact, as demonstrated in [18], all the pairwise dot products between \( x'_{u,m} \) and other subsequences are included in \( Q_u \), and they can be efficiently retrieved as

\[
M_{u,i} = Q_{u,m-1+i}, \quad \text{for } 1 \leq i \leq N - m + 1
\]

As can be observed in (10)-(14), by performing sequential dot products once and DFT based calculation three times, the original estimation of \((N-m+1)\) sequential dot-products can be quickly finished in a vectorized way. This would lead to a significant acceleration for searching STNNs. Moreover, based on the recursive relationship between successive dot products, further speed-up can be achieved. Given the dot product value \( M_{u,v} \), one can easily estimate \( M_{u+1,v+1} \) as

\[
M_{u+1,v+1} = M_{u,v} - x'_u x'_v + x'_{u+m} x'_{v+m}
\]

IV. Simulation Test in Benchmark System

The proposed approach was first tested on the Nordic test system to verify its efficacy. This is a benchmark system simplified from the actual Swedish and Nordic power grid [20]–[22]. As shown in the shaded area of Fig. 2, the five adjacent 130-kV buses in its receiving-end area, i.e., buses 1041~1045, were assumed to be configured with PMUs for online monitoring. Supposing that the system encountered transient events such as three-phase short circuits, time-domain simulations were conducted to simulate the acquisition of sequential PMU data during system dynamics. Specifically, PMU data were acquired with a sampling rate of 100 Hz. The length of the OTW was set to 5 s. With 500 data points in the OTW, the length of subsequences for STNN discovery was empirically specified as 1/10 of the OTW, i.e., \( m = 50 \).

Taking voltage measurements for example, the reliability and efficiency of the proposed approach for BPDD in various typical scenarios are demonstrated below.

A. BPDD in Data Spike Scenario

A relatively small data spike was imposed on the voltage profile of bus 1041, as shown in Fig. 3(a). During STNN
temporal anomalies and the corresponding subsequences with low-level noises. Specifically, Gaussian white noises with the signal-to-noise-ratio set to 40 dB were partially superposed on the system indeed help to distinguish abnormal data sensing errors from its natural dynamics. The proposed approach not only accurately detects the small data spike, but also bypasses the much more significant voltage deviation caused by the transient event is experienced in the same OTW. In particular, starting from the peak point with five steps from the peak point are examined using (8). In this way, all the potential anomalies are screened out, as depicted in Fig. 3(b). The corresponding subsequences are further characterized in Fig. 3(c).

Clearly, the data spike is successfully detected by the proposed approach. In fact, as can be observed in Fig. 3, compared with the slight data spike, a more significant voltage deviation caused by the transient event is experienced in the same OTW. The proposed approach not only accurately detects the small data spike, but also bypasses the much more fluctuating transient event data. Such a strong discriminability reveals that the inherent spatial-temporal correlations within the system indeed help to distinguish abnormal data sensing errors from its natural dynamics.

B. BPDD in Highly Noisy Scenario

The proposed approach was further tested in the presence of high-level noises. Specifically, Gaussian white noises with the signal-to-noise-ratio set to 40 dB were partially superposed on the voltage profile of bus 1041, as depicted in Fig. 4(a). Following the similar BPDD procedure presented above, spatial-temporal anomalies and the corresponding subsequences with bad data were figured out, as shown in Fig. 4(b)-(c). Obviously, the bad data with high-level noises are correctly identified as spatial-temporal anomalies. Analogous to the above BPDD results, the proposed approach avoids false detection of the event data.

C. BPDD in Unchanged Data Scenario

Assuming the potential transformer (PT) configured for the PMU at bus 1041 temporarily encounters device failures from 2.5 s to 2.7 s, the corresponding voltage measurements of bus 1041 become unchanged during the time interval. Given this scenario, whether the proposed approach can identify such abnormal measurements was tested here. As exhibited in Fig. 5, the proposed approach accurately detects the unchanged subsequences as well, without committing false alarm on the event data.

D. Comprehensive BPDD Performances

Taking into account typical operating conditions, topological changes and fault locations in the Nordic test system, 600 dynamic cases subject to various contingencies were generated by time-domain simulations. In each case, the above three types of bad data were imposed onto the voltage profile of bus 1041 to simulate anomalies. The OTW sliding with a step of $\Delta T = 1$ s was utilized to extract five windows of PMU data from buses 1041~1045 for each case. Based on these settings, 9000 PMU measurement matrices were obtained for BPDD. For comparative study, the representative LOF based method in [15] was also carried out. To make the comparison reasonable, the two LOF thresholds in [15] were carefully tuned to achieve the optimal performance. Both
E. Computational Efficiency of BPDD

The computational efficiency of the BPDD approach was tested here by recording its computation time in each scenario. All the tests were carried out using a PC configured with a 3.60-GHz×8 Intel Core i7-7700 CPU and 32.0 GB RAM. For comparative study, a brute-force method without adoption of the fast STNN discovery strategy (see Section III-B) is conducted for BPDD as well. The time consumptions of the two methods in the three scenarios are summarized in Table II. Evidently, after the adoption of the fast STNN discovery strategy, the computational efficiency of BPDD is dramatically improved by more than 10 times. Concretely, it costs the proposed approach less than 0.35 s to complete BPDD for a OTW of 5 s. In practical online monitoring, highly efficient streaming BPDD can be performed by continuously sliding the OTW with a time step of 0.4∼0.5 s.

V. Experimental Test in Practical System

With field PMU data collected from the real-world CSG, the proposed approach was further tested in practical contexts to show its applicability and scalability. In particular, synchronous voltage measurement matrices with a OTW of 16 s were acquired from CSG for case study here. The PMU sampling rate was specified as 25 Hz in CSG. Taking a small region in Guangdong Province with seven 500-kV buses for example, i.e., the Guangzhou subsystem depicted in Fig. 6 [23], the seven buses’ voltage profiles were collected for BPDD tests. Note that, unlike the simulation based PMU measurements in test systems, as there exist multiple channels to measure bus voltages at a single substation (bus), multiple voltage TS can be obtained for the same bus in the bulk system. Considering this practical situation, a two-layer BPDD scheme was designed here: 1) At each substation, multi-channel PMU data are gathered and concatenated as a spatial-temporal TS for BPDD; those channels identified to be contaminated with bad data are excluded from the data sources, and the substation’s voltage profile is estimated by averaging the voltage sequences of the remaining channels. 2) In the specific region, all the substations’ voltage profiles are collected and concatenated to perform BPDD, which is similar to that in the Nordic test system.

A. BPDD in Substation Layer

Without loss of generality, a substation was randomly chosen from the region to illustrate the performance of BPDD in the substation layer. As presented in Fig. 7, there are eight channels at this substation. However, the voltage measurements in the seventh channel remain unchanged with zero value in the whole OTW. In addition, there exist three fairly small data spikes (about 0.005 pu) in the fourth channel. All of these measurements are actually bad data at the substation. With 400 data points in the 16-sec OTW, the length of subsequence was simply set to \( m = n/10 = 40 \) for STNN discovery. Fig. 8 summarizes the BPDD results.

As can be seen, the proposed approach successfully identifies all the bad data in the substation. Not only all the abnormal zero-value PMU measurements are effectively detected, but also the three extremely weak data spikes are correctly recognized. Besides, it is noticed that no false detection is made on the event data, although they have a much larger fluctuation of 0.2 pu than the data spikes. Note that such a highly
Fig. 6. Structure of China Southern Power Grid [23].

Fig. 7. Multi-channel voltage profiles at a substation.

Fig. 8. BPDD with multi-channel voltage TS.

accurate BPDD is finished in 0.413 s, which well satisfies the online requirement on high efficiency. As multiple substations’ measurements need to be examined, parallel computing can be carried out to further accelerate BPDD in the substation layer.

B. BPDD in Regional Layer

After the completion of BPDD in the substation layer, all the seven substations’ PMU measurements were integrated as a TS profile to perform regional BPDD. The corresponding BPDD result is shown in Fig. 9. As can be seen, all the bad data in the form of an abnormally unchanged voltage value (0.454 pu) in the entire OTW (collected from one substation) was accurately identified by the proposed approach. The reason why the anomalous data are not detected by the substation layer’s BPDD lies in that all the channels’ voltage measurements at the substation remain unchanged. Essentially, they do not go against the inherent spatial-temporal correlations between multiple channels at the substation. Nevertheless, based on the spatial-temporal correlations between adjacent substations, these bad data are successfully filtered out in the regional layer. The proposed approach spends 0.386 s to achieve BPDD in the regional layer. Hence, the whole BPDD task can be completed in no more than 0.8 s, resulting in a high online efficiency.

C. Comprehensive BPDD Performances

With nearly 2 hours of field PMU measurements collected from CSG in August, 2018, the 16-sec OTW sliding with a step of $\Delta T = 1$ s was utilized to acquire 6000 realistic synchronous measurement matrices. Based on the two-layer BPDD scheme, voltage measurements at each substation were first filtered out by the proposed approach, and they were then averaged for BPDD in the regional layer. For regional-layer BPDD, similar to the tests on the Nordic test system, both the proposed approach and the LOF based method [15] were performed for comparative study. Taking a certain substation for instance, its comprehensive BPDD results are summarized in Table III. Again, the two methods achieve comparable BPDD performances. However, it should be pointed out the preparation of a clean learning database for LOF based method is extremely time-consuming and costly in practice, because one has to resort to system operators with special domain expertise for class labeling. In this respect, the proposed
approach without such requirements would be preferred in practical online monitoring.

VI. CONCLUSION

Based on the inherent spatial-temporal correlations during power system dynamics, this paper develops a model-free TS data-driven approach for online BPDD. With no need for labeling bad PMU data in advance for offline learning, it performs unsupervised BPDD in a a highly efficient way. Specifically, following the idea that spatial-temporal anomalies are significantly different from their STNN, sequential BPDD is carried out by performing fast STNN discovery and filtering out those subsequences with abnormal STNN distance values. With no requirement on iterative learning, it gets rid of time-consuming offline learning, being suitable for handling online PMU data streams. Numerical test results on the Nordic test system show that the proposed approach achieves excellent performances in various bad data scenarios. Further tests with field PMU data in CSG demonstrate the scalability of the BPDD approach in practical contexts. In relevant future work, by combining explicit domain knowledge with machine learning techniques, semi-supervised learning schemes would be designed to further improve the overall accuracy of BPDD.

REFERENCES

[1] G. Bedi, G. K. Venayagamoorthy, R. Singh, R. R. Brooks, and K.-C. Wang, “Review of internet of things (iot) in electric power and energy systems,” IEEE Internet of Things Journal, vol. 5, no. 2, pp. 847–870, 2018.
[2] J. Yu, D. J. Hill, V. O. Li, and Y. Hou, “Synchrophasor recovery and prediction: A graph-based deep learning approach,” IEEE Internet of Things Journal, vol. 6, no. 5, pp. 7348–7359, 2019.
[3] Y. Xu, R. Zhang, J. Zhao et al., “Assessing short-term voltage stability of electric power systems by a hierarchical intelligent system,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 8, pp. 1686–1696, 2016.
[4] Y. Ge, A. J. Flueck, D.-K. Kim et al., “Power system real-time event detection and associated data archival reduction based on synchrophasors,” IEEE Trans. Smart Grid, vol. 6, no. 4, pp. 2088–2097, 2015.
[5] I. Kamwa, S. Samantaray, and G. Joos, “Compliance analysis of pmu algorithms and devices for wide-area stabilizing control of large power systems,” IEEE Trans. Power Syst., vol. 28, no. 2, pp. 1766–1778, 2012.
[6] G. N. Korres and N. M. Manousakis, “State estimation and bad data processing for systems including pmu and scada measurements,” Electric Power Systems Research, vol. 81, no. 7, pp. 1514–1524, 2011.
[7] S. G. Ghiocei, J. H. Chow, G. Stefopoulos et al., “Phasor-measurement-based state estimation for synchrophasor data quality improvement and power transfer interface monitoring,” IEEE Trans. Power Syst., vol. 29, no. 2, pp. 881–888, 2014.
[8] Y. Lin and A. Abur, “A highly efficient bad data identification approach for very large scale power systems,” IEEE Trans. Power Syst., vol. 31, no. 6, pp. 5979–5989, 2016.
[9] K. D. Jones, A. Pal, and J. S. Thorp, “Methodology for performing synchrophasor data conditioning and validation,” IEEE Trans. Power Syst., vol. 30, no. 3, pp. 1121–1130, 2014.
[10] W.-T. Li, C.-K. Wen, J.-C. Chen, K.-K. Wong, J.-H. Teng, and C. Yuen, “Location identification of power line outages using pmu measurements with bad data,” IEEE Trans. Power Syst., vol. 31, no. 5, pp. 3624–3635, 2016.
[11] B. Gou and R. G. Kavasseri, “Unified pmu placement for observability and bad data detection in state estimation,” IEEE Trans. Power Syst., vol. 29, no. 6, pp. 2573–2580, 2014.
[12] P. Gao, M. Wang, S. G. Ghiocei et al., “Missing data recovery by exploiting low-dimensionality in power system synchrophasor measurements,” IEEE Trans. Power Syst., vol. 31, no. 2, pp. 1006–1013, 2016.
[13] Y. Hao, M. Wang, J. H. Chow et al., “Modelless data quality improvement of streaming synchrophasor measurements by exploiting the low-rank hankel structure,” IEEE Trans. Power Syst., vol. 33, no. 6, pp. 6966–6977, 2018.
[14] M. Wu and L. Xie, “Online identification of bad synchrophasor measurements via spatio-temporal correlations,” in Power Systems Computation Conference (PSCC). IEEE, 2016, pp. 1–7.
[15] M. Wu and L. Xie, “Online detection of low-quality synchrophasor measurements: A data-driven approach,” IEEE Trans. Power Syst., vol. 32, no. 4, pp. 2817–2827, 2017.
[16] A. Mueen, S. Nath, and J. Liu, “Fast approximate correlation for massive time-series data,” in Proc. ACM SIGMOD International Conference on Management of Data. ACM, 2010, pp. 171–182.
[17] J. Zakaria, A. Mueen, and E. Keogh, “Clustering time series using unsupervised-shaplets,” in IEEE International Conference on Data Mining (ICDM). IEEE, 2012, pp. 785–794.
[18] C.-C. M. Yeh, Y. Zhu, L. Ulanova et al., “Matrix profile ii: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets,” in IEEE International Conference on Data Mining (ICDM). IEEE, 2016, pp. 1317–1322.
[19] Y. Zhu, Z. Zimmermann, N. S. Senobari et al., “Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins,” in IEEE International Conference on Data Mining (ICDM). IEEE, 2016, pp. 739–748.
[20] T. Van Cutsem and L. Papangelis, “Description, modeling and simulation results of a test system for voltage stability analysis,” Université de Liège, Tech. Rep., 2013.
[21] L. D. P. Ospina, A. F. Correa, and G. Lammert, “Implementation and validation of the nordic test system in digsylent powerfactory,” in 2017 IEEE Manchester PowerTech. IEEE, 2017, pp. 1–6.
[22] L. Zhu, C. Lu, Z. Y. Dong, and C. Hong, “Imbalance learning machine-based power system short-term voltage stability assessment,” IEEE Trans. Ind. Informat., vol. 13, no. 5, pp. 2533–2543, 2017.
[23] L. Zhu, C. Lu, and Y. Luo, “Time series data-driven batch assessment of power system short-term voltage security,” IEEE Trans. Ind. Informat., Early Access, 2020.