Denoising and Bad Data Detection in Distribution Phasor Measurements using Filtering, Clustering and Koopman Mode Analysis

Amirkhosro Vosughi, Amir Gholami, Anurag K. Srivastava, *Senior Member, IEEE*

***Abstract*—Distribution-level phasor measurement units (D-PMU) data are prone to different types of anomalies given complex data flow and processing infrastructure in an active power distribution system with enhanced digital automation. It is essential to pre-process the data before being used by critical applications for situational awareness and control. In this work, two approaches for detection of data anomalies are introduced for offline (larger data processing window) and online (shorter data processing window) applications. A margin-based maximum likelihood estimator (MB-MLE) method is developed to detect anomalies by integrating the results of different base detectors including Hampel filter, Quartile detector and DBSCAN. A smoothing wavelet denoising method is used to remove high-frequency noises. The processed data with offline analysis is used to fit a model to the underlying dynamics of synchrophasor data using Koopman Mode Analysis, which is subsequently employed for online denoising and bad data detection (BDD) using Kalman Filter (KF). The parameters of the KF are adjusted adaptively based on similarity to the training data set for model fitting purposes. Developed techniques have been validated for the modified IEEE test system with multiple D-PMUs, modeled and simulated in real-time for different case scenarios using the OPAL-RT Hardware-In-the-Loop (HIL) Simulator.**

***Index Terms*—D-PMU, Active Distribution Network, Kalman Filter, Koopman Mode Analysis, Data Anomaly Detection, DB-SCAN, Margin-based Maximum Likelihood Estimator, Measurement Denoising.**

I. Introduction

A. Motivation

Transforming from an inactive to an active distribution network with the integration of DERs and microgrid, it is essential to utilize new data sources to increase the observability of the system to enable centralized coordination between different agents in the smart distribution grid. Synchrophasor data provides a wealth of information that enables the grid operator to capture fast transient dynamic events. However, D-PMU data is needed to be pre-processed in order to denoise and detect bad data before being fed to any other critical applications.

Power system is a multi-layer distributed cyber-physical system. Synchrophasor data Anomalies may originate in different cyber-physical system layers. Fig. 1 illustrates data flow of synchrophasor data in different layers of smart grid. Spatial-temporal correlated anomaly originated from physical layer and related to physical events. In sensor layer, D-PMU data is generated with contaminated data by Bad Data and noise from base sensors (voltage/current/GPS). Also, this data needs to be sent over the communication network and it is prone to communication delay, failure, and cyber attacks. Anomalies in sensor and communication layer, are not usually temporal and spatial correlated. Available data in application layer is contaminated with anomalies of previous layers and distinguishing between the source of the anomaly is challenging. Fig. 2 determines different types of events that can originate anomaly in D-PMU data. In general, it is hard to differentiate between bad data anomaly and event-based anomaly. This leads to confusion between the origin of anomalies including but not limited to noise and BDD and event, sensor problems, cyber attack, and physical events. This must be done based on Spatial-temporal correlation using Machine Learning (ML) and a data-driven approach. Non-Spatial-temporal anomaly must be compensated (denoising and replacing bad data); however, the signature of Spatial-temporal anomaly events needs to be preserved since they provide useful data about the dynamic event in the system. Model-based approaches to BDD and denoising are not practical for high sample rate D-PMU data. Furthermore, Power system is subjected to several frequent changes due to the contribution of independent agents, topology reconfiguration, and stochastic nature of new energy resources that make use of model-based approaches even more challenging.

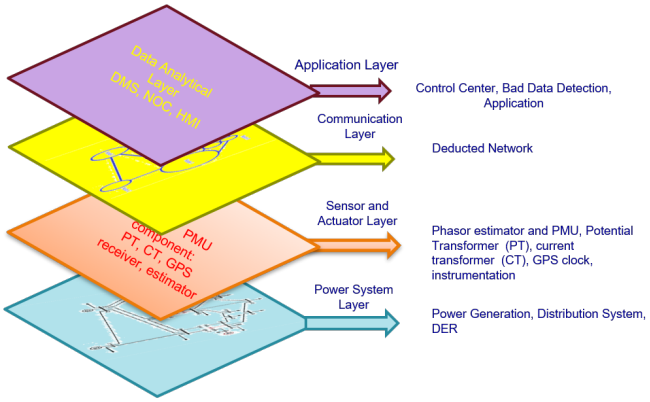


Fig. 1: Synchrophasor data in different layers of power distribution system

A decision tree can be used for the identification and

The authors are with the School of Electrical Engineering and Computer Science at Washington State University. This work has been partially supported by the US Department of Energy DE-EE0008775 SolarSTARTS and DE-IA0000025 UI-ASSIST.

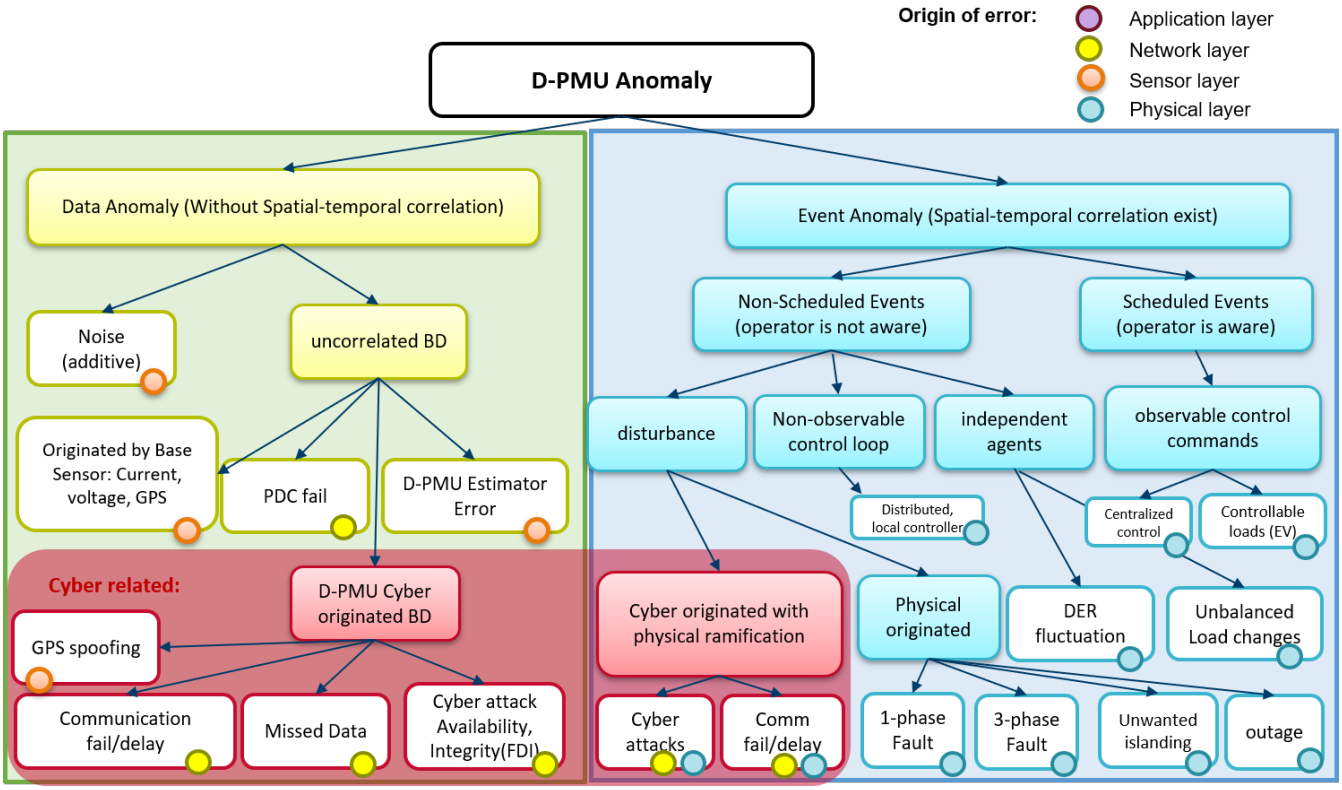


Fig. 2: Possible Causes for D-PMU Data Anomalies

handling of D-PMU Anomaly. The denoising technique can be used to remove high-frequency noise from the time series. Non-inertia anomaly must be evaluated based on temporal correlation with previous/next measurement and spatial correlation with other measurements to distinguish between bad data and event-induced disturbances. After detection of missed data and non-inertial anomalies can be replaced with appropriate values based on an educated guess with an interpolation or prediction technique. Subsequently, the event can be analyzed further for localization and classification; however, if the effect of the event is not well-correlated with other measurements, the hypothesis of associating the anomaly to the dynamic event can be revisited. Mis-classification between bad data (BD) and special-temporal correlated events could have serious consequences. The False positive in detecting non-Spatial-temporal BD, result in losing signature of event and weaken the credence of classification/localization ML and data-driven methods, in case that signal is used in the close loop, the controller becomes less responsive. False-negative in detecting non-Spatial-temporal BD, result in associating BD to the physical event which misleads the classification/localization in ML and data-driven methods, in the case that signal is used in the close loop, it causes big measurement disturbance.

B. Related Works

Data pre-processing and measurement denoising is a significant step which needs to be accomplished prior to investigation of any physical or cyber-induced events throughout the system [1]. A proper BDD technique is of enormous concern in such Distribution System (DS) applications as State Estimation (SE) [2], fault classification [3], fault location identification [4], and situational awareness [5]. [6] presents a comprehensive study on different ML-based approaches for distribution level disturbance analysis with regard to efficacy examination and performance evaluation of each technique. Assuming a normal distribution can be estimated for load profiles, authors of [7] propose a novel BDD approach based on the Weighted Least Squate (WLS) method. Literature is mainly focusing on either BDD or denoising separately accompanied with detection of sensor faults in applications with renewable energies [8], while in the current research work, the focus is on development of two tools for online and offline BDD and denoising together.

Wavelet transform has been used widely for transient analysis in the power system [9], [10], [11] and is employed for signal smoothing in this work. We combine the results from a variety of base anomaly detectors such as Hampel filter [12], Quartile technique, and DBSCAN using MB-MLE. For online signal processing, the recursive Bayesian Kalman filter (KF) has been used [13]. For the prediction step of KF, rather than using a physical model, we utilize Koopman Mode Analysis (KMA) [14] for fitting the model for prediction of behaviour of the system with a data-driven approach. A lot of application for KMA in power system has been developed including but not limited to power system stability and dynamic analysis [15], control and estimation [16], and fault line isolation [17] and attack detection [18]. There has been other studies of voltage stability-based analysis incorporating HVDC lines, including cyber and physical co-simulation in [19] which are also of

high potential fits for KMA applications.

C. Contributions

The focus of this work is to detect anomalies in D-PMU data and denoising of the measurements. Based on the application, it might need to be done in online manner for close loop control and monitoring purpose or offline in mode for further investigation of the nature of the events. Here, we develop two approaches for dealing with each case. For offline cases, the computational time cost is relaxed and bad data detection and denoising can be done with smoothing methods. For that, MB-MLE is used to detect the bad data by integrating the base detector scores. Different base anomaly detectors are employed such as Hample filter, Quartile-based outlier detection, and DBSCAN, of various methodology types to detect the anomalies and this diversity enhances the likelihood of isolating outliers. The wavelet smoothing technique is employed to denoising the signal. For online mode, the KF has been used to denoise the signal and detect the bad data based on the residual analysis. Since, it is hard to obtain observability towards the dynamic of distribution network from high sample rate synchrophasor data, data-driven Koopman mode analysis has been used to fit a model that describes the underlying interaction dynamic between the D-PMU measurements which is employed in the prediction step of Kalman Filter (KF). For preparing data to fit a model, denoising and outlier detection techniques developed in the offline phase are used for pre-processing. Since the set of data has been used for the trained model, might not be comprehensive, the online evaluation of the performance of the fitted model is essential and the parameters of the KF must be adjusted adaptively based on the credence of the fitted model to rely more on measurement as necessary. This is done by observing the deviation of the mean of the estimated state from measurements.

The rest of this paper is outlined as follows. In section 2, the offline techniques for anomaly detection and denoising of the system is introduced. Section 3 presents our approach for online data-driven anomaly detection and denoising using data-driven adaptive KF. In section 4, the results of the implementation of algorithms on real-time simulator data are provided. Fig. 3 illustrates the connection of proposed approaches.

To summarize, the main contribution of this paper can be enumerated as follows:

1) Development of an offline tool with larger time window data processing in order to achieve a high precision denoising (wavelet-based) and D-PMU BDD using margin-based maximum likelihood estimator (MB-MLE) for applications with less time sensitivity
2) Development of an online tool with shorter time window data processing for rapid denoising and BDD of D-PMU continuous data streams for fast responding close loops and online monitoring control room applications based on an adaptive KF
3) Extension of KF implementation in feeders with sparse high sampling rate D-PMU measurements, without any knowledge of system physical model, as a result of Koopman Mode Analysis (KMA) employment

II. Offline tool for D-PMU anomaly detection

A. Wavelet Filter

Wavelet transform is an extension of Fourier transform where frequency and temporal analysis is performed simultaneously to capture signal frequency evolution through time. For that, wavelet transform localize features in the data set with different scales. The fundamental principle of wavelet transform is the sparse representation of the temporal signal that means the signal can be presented in the limited large-magnitude coefficient and the small value wavelet coefficient are related to noise which can be removed without affecting the original signal of interest. After thresholding, the signal is reconstructed with inverse wavelet transform. In this work, Wavelet transform is utilized to D-PMU signal denoising.

B. Hampel Filter (HF)

Hample filter evaluates data in a sliding window of $2n'$ neighbor samples with n' in each side. The median absolute deviation (MAD) is computed as follows:

$$MAD = Median(|\bar{X} - Median(X)|) \tag{1}$$

where $\bar{X}$ presents the window in which the central element is the sample of interest x and $Median(.)$ calculate the local median in the input array. MAD provides an estimation of standard deviation on the window $\sigma = 1.4826MAD$.

If the difference of sample from the median of the window, it bigger than three times of estimated standard deviation σ, It is considered an anomaly as a convention for Hampel filter. For obtaining the margin of BD, the distance on the sample from the median is computed as $\delta x = x - Median(X)$. The logarithm of dividing this value to three times of estimated standard deviation gives a metric about the margin of the possibility of being an outlier. If this metric is positive, the sample is considered an outlier and vice versa. In the case that metric is positive, The bigger the metric, the more possibility to be an outlier from the Hampel filter perspective. If the value is close to zero, the Hample filter is mostly indecisive about the sample. Hampel is prone to detect a false bad data anomaly in steady-state since the standard deviation in the window is fairly small and a tiny deviation from the median can be considered as a bad data anomaly. In that case, replacing with median will not be harmful. However, since we are interested to detect anomalies with different approaches and combine the results based MB-MLE, we define the small constant ϵ_h to reduce the chance of detecting outlier in steady-state. Therefore margin bad data HF defines for each sample as

$$Margin_{HF} = \log(\frac{\delta x}{3\sigma + \epsilon_h}) \tag{2}$$

C. Quartile-based Anomaly Detection (QB)

Another statistical approach for anomaly detection is quartile-based anomaly detection. The advantage of this method is the fact that it does not have a normal distribution assumption of the data. To implement, data is analyzed in small windows, and the median and quarter of the window

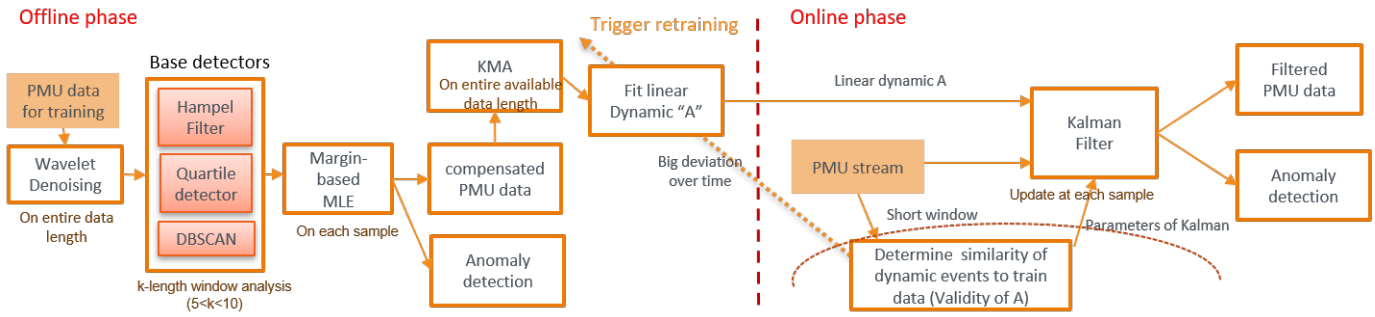


Fig. 3: The scheme of proposed online and offline approaches for D-PMU non spatial temporal anomaly detection

are computed. The sample is marked as an anomaly if they are 1.5 times bigger than the upper quartile or if they 1.5 times below the lower quartile. For obtaining the margin metric of being an outlier, the distance of each sample from the median is computed (δx_i) and divided by the distance of the corresponding quartile from the median (δq). The logarithm of this division gives insight into how likely the sample is an outlier.

Like Hampel filter, this approach is prone to detect false positives in steady-state. By the same token, the constant constant ϵ_q is added to the denominator of the fraction. Therefore, margin quartile method can be computed for each sample as

$$Margin_{QB} = \log(\frac{\delta x_i}{\delta q + \epsilon_q}) \tag{3}$$

D. Density-based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN has been deployed for determining outliers in finding the disturbances from the output of an unsupervised model [20]. DBSCAN uses two tunable parameters include threshold ϵ and Minimum Number of Points (MinPts). The data point in ϵ radius of cluster is merged to it. At least MinPts data points are needed to form a cluster. The data points without clusters are determined as outliers.

In this work, to find the margin-based value, the parameter threshold ϵ which makes each measurement categorized as an outlier is calculated based using grid search. Subsequently the parameter $Margin_{DBSCAN} = \log \frac{\epsilon}{\epsilon_0}$ is found to determine the margin of begin outlier with the DBSCAN method for each sample. The value of ϵ_0 can be estimated based on results from ground truth data.

E. Margin-Based Maximum Likelihood Estimator (MB-MLE)

Different calculated margin from base-detectors are combined together to make a robust decision about the likelihood of being bad data anomaly for each sample as follows:

$$MLE = \mathcal{N}(\sum_{i \in \mathfrak{B}} c_i Margin_i) \tag{4}$$

where set $\mathfrak{B} = \{HF, QB, DBSCAN\}$ show the available based detectors and c_i is associate confidence factor which can be obtain based on historical data in the ground truth is know or using majority voting to estimate the ground truth. For this work, we assume $c_i = 1$ for all $i \in \mathfrak{B}$. Positive MLE indicates anomaly and vice versa and the biggest deviation from zero, indicates more confidence of MB-MLE bad data anomaly detector.

III. Online tool for D-PMU anomaly detection

Although the introduced offline tools in the previous section, can determine the anomalies and reduce noise effect with smoothing techniques with good accuracy. The implementation of those techniques is time-consuming and they might not be suitable for the application when fast decision making is needed. Considering these facts, in this section, we develop an online D-PMU pre-processing tool using Koopman Mode Analysis and adaptive Kalman Filter by adjusting the KF parameters based on the confidence on of the fitted model with online monitoring. This technique used both temporal and spatial correlation between synchrophasor measurement to signal processing.

A. Kalman Filter

KF is an optimal recursive bayesian estimator. If the assumptions of the KF (Linear Dynamic, Gaussian process and measurement noise) hold, it surpasses any other casual filter. It consists of two steps namely prediction and correction. In the prediction step, the model of the system is used to predict the expected value and uncertainty of the next step measurement.

$$\begin{aligned} \hat{x}_{n|n-1} &= A\hat{x}_{n-1|n-1} + Bu_n \\ P_{n|n-1} &= AP_{n-1|n-1}A' + Q_n \end{aligned} \tag{5}$$

Where A is the linear dynamic of the system, B is signature of input, and u_n in control input at time n.$\hat{x}_{n-1|n-1}$ and $P_{n-1|n-1}$ shows the prior believe on the expected value and covariance of estimated states. $\hat{x}_{n|n-1}$ and $P_{n|n-1}$ represents those at time n based on prediction and Q_n is process noise covariance.

In the correction step, the value of prediction is modified based on new observation, and the uncertainty of perdition reduced based on characteristics of observation.

$$\begin{aligned} S_n &= CP_{n|n-1}C' + R_n \\ K_n &= P_{n|n-1}C'S_k{}^{-1} \\ \hat{x}_{n|n} &= \hat{x}_{n|n-1} + K_n(y_n - C\hat{x}_{n|n-1}) \\ P_{n|n} &= P_{n|n-1}(I - K_nC) \end{aligned} \tag{6}$$

Where C is signature of output. R_n is covariance of measurement noise and S_n characterize the uncertainty of measurement. K_n is Kalman gain, y_n is measurement at time n and $\hat{x}_{n|n}$ and $P_{n|n}$ shows the posterior believe on the expected value and covariance of estimated states after observation.

In our formulation, y is direct and indirect measurements from different D-PMU contaminated with noise and bad data anomaly and x indicates measurement without contamination and $\hat{x}$ is estimation of x, B is equal to zero and C equal to identity matrix with appropriate size therefore y has same size as x. In the next subsection, we explain how to obtain linear dynamic model A and adjust parameters R_n and Q_n attentively.

B. Koopman Mode Analysis

Koopman Theory asserts that any underlying nonlinear system can be described completely with infinite state space [14]. To find a finite approximation of the Koopman states, many data-driven mechanisms are suggested. In this work, we use Dynamic Mode Decomposition to this end. DMD is a data-driven approach for fitting a linear dynamic to a set of measurements [21], [22]. Consider that the matrix $Z_1^{k'} = [z_1, ..., z_{k'}]$ represents a window of k' snapshots consisting n measurement at each snapshot where vector z_i shows the ith snapshot. The matrix $Z_1^{k'}$ is constructed by stacking different D-PMU's direct measurement data streams such as voltage/current magnitude/angel, and frequency and indirect measurement data stream including active/reactive power flow on each other. The notation Z_p^q shows a set of subsequent data where subscript p is the index of the starting snapshot and superscript q is the index of the last snapshot in the window. We assume the sampling distance between every two successive snapshots is constant and is shown with δ. Suppose there exist a linear dynamic A maps data sample z_i to data sample z_{i+1} for $i = 1, \cdots, k' - 1$ as

$$z_{i+1} = Az_i \tag{7}$$

Generally speaking, measurements are generated from a nonlinear underlying dynamic. Using DMD, we attempt to find the best linear dynamic approximation that describes the relationship between the measurements over the window. In other words, we try to minimize the residual of the linear system defines as

$$r = \left\| Z_2^{k'} - AZ_1^{k'-1} \right\|_2 \tag{8}$$

Where $\|.\|_2$ shows norm 2 of matrix. A naive approach to compute linear system is $A = Z_2^{k'} Z_1^{k'-1^+}$ where $Z_1^{k'-1^+}$ is Moore-Penrose pseudo-inverse of $Z_1^{k'-1}$. The matrix inverse can be obtained via the LQ method. Nonetheless, Singular Value Decomposition (SVD) allows more robust numerical stability [23]. However, it is not computationally efficient to calculate A directly because k' might be large. Therefore, the DMD algorithm finds eigenstructure linear dynamic without computing it directly. By employing SVD we have

$$Z_1^{k'-1} = U\Sigma V^* \tag{9}$$

Regardless of the possible hug dimensions of A, in most cases, the underlying dynamic of the window can be delineated with a few m dominant modes which are of the notable portion of event energy. Using these modes, the equivalent linear dynamic $\tilde{A}$ is computed as follows

$$\tilde{A} = U_m^* Z_2^{k'-1} V\Sigma^{-1} \tag{10}$$

Where U_m shows first m columns of U. $\tilde{A}$ has same eigenvalues as A. By eigenvalue decomposition of $\tilde{A}$ we have $\tilde{A}W = \Lambda W$ where diagonal matrix Λ contains eigenvalues of A and $\tilde{A}$ matrices. Columns of W are eigenvectors of $\tilde{A}$. Eigenvectors of A are computed as $\Psi = V_2^{k'} V\Sigma^{-1}W$. The jth diagonal entity of Λ indicates eigenvalue of jth mode shown by λ_j and corresponding eigenvector is jth column of matrix Ψ denoted by ψ_j.

To Compute energy amplitude of each mode, SVD of the data window is obtained as $Z_1^{k'} = U_0\Sigma_0 V_0^*$. The matrix $\Lambda_t = \exp([\lambda_{1,t}, \cdots, \lambda_{m,t}]^T)$ indicates evolution of each mode in the window where $\lambda_{i,t} = \lambda_j \bar{T}$ and $\bar{T} = [0, h, 2h, \cdots, (k'-1)h]^T$. Subsequently, we calculate matrix $\Xi_1 = \Upsilon^T\Upsilon \odot (\Lambda_t\Lambda_t^T)^*$ where $\odot$ denote Hadamard product operator, superscript $*$ indicates Hermitian transpose operator and $\Upsilon = U_0^T\Psi$. Further, $\Xi_2 = Diag(\Lambda_t V_0\Sigma_0\Upsilon)^*$ where the operator $Diag(.)$ place the diagonal entries of square input matrix into a vertical vector. Finally, m-length vector $\Xi = \Xi_1^{-1}\Xi_2$ is computed where absolute value of the jth element of Ξ is equal to σ_j which shows energy amplitude jth mode of linear system.

Here, we use KMA for finding linear system A that describes the dynamics of the system based on long and comprehensive historical data after pre-processing with offline anomaly and denoising techniques that were introduced in section 2. From each D-PMU direct measurements such as voltage/current magnitude/angle and frequency and indirect measurements such as active and reactive power flow in different phases are used to find the linear model. The reason for adding indirect measurement is that it is proofed by adding more observant, the better linear system can be fitted to describe the underlying dynamic of system [21]. The fitted linear model is used for prediction in online KF for denoising and bad data detection.

C. Adaptive Adjustment of Kalman Filter Parameter

With the assumption of having Gaussian Noise, the mean of estimated value and measurement value overtime must have the same values. This can be used as an indicator that the fitted model A, does not describe the dynamic to the system precisely. The deviation of mean in all estimated

measurements from noisy measurement over a is calculated as $N[n]$ in each sample n. The average of that over that in short interval on recent samples with length 5 is called $\bar{N}$. This metric is used to check the credence of the fitted linear model. Using this parameter, the metric ρ is defined as

$$\rho = \frac{2\sigma_{max}}{\exp\{-\sigma_{steep}\bar{N}\} + 1} - \sigma_{max} + 1 \tag{11}$$

In the above function, ρ goes to 1 as when N is zero and it goes to $\sigma_{max} + 1$ as N goes to ∞. σ_{steep} determines how fast this transition occurs. The process noise of the KF is then updated attentively as $Q = \rho Q_{base}$. It needs to be clarified, that A is computed offline and used in online pre-processing using KF. However, if ρ remains bigger than a preset threshold for a long period of time (1 day), that means that power grid topology or dynamic has been changed significantly and a new linear model A required to be fitted to obtain good results with online tool.

D. Online model update scheme

If the value of parameter N stays big for a considerable amount of time, it means that the linear model A cannot describe the behavior of the system anymore and it is essential to retrain the model. This will trigger offline analysis automatically to compute a new linear dynamic A. For that, we compute the parameter $\tilde{N} = \sum_{k=n-n_0}^{n} N[k]$ where $[n-n_0, n]$ indicates the time window that we accumulate the average deviation of mean in all estimated measurement from noisy measurement over. n_0 is chosen to have a time window that covers half and hour and the model is retrained if the $\tilde{N}$ if bigger that the threshold γ.

IV. Performance Evaluation

A. Testbed and Test Case Development

For the evaluation purposes of the developed algorithms in this work, IEEE 33-node test feeder is chosen to be modeled and simulated in Opal-RT hardware-in-the-loop (HIL) simulator.

Fig. 4 shows the single-line diagram of the developed IEEE-33 node system in OPAL-RT with extra modifications compared to the standard model. The HYPERSIM interfacing simulator is used for software aspect of the model development with capabilities of DER integration and D-PMU simulation. The model is capable of being executed in real-time with 50 μs timestep and with hardware D-PMUs added to the loop of the simulation. The developed Battery Energy Storage Systems (BESS) and PV units are modeled to be operating under the grid-forming configuration, making the distribution feeder to be able to operate under a 100% renewable scenarios in an islanded operation. The modifications on the standard IEEE system include adding four BESS and PV combined energy resources at four different loaded nodes of the feeder. The network has three radial feeders with possible mesh interconnection among themselves. Simulation D-PMUs are modeled and placed at generation nodes: $\{1, 14, 18, 22, 29, 33\}$. The monitored instantaneous time domain signals captured from the OPAL-RT solver are sent to simulated D-PMUs for phasor estimations.

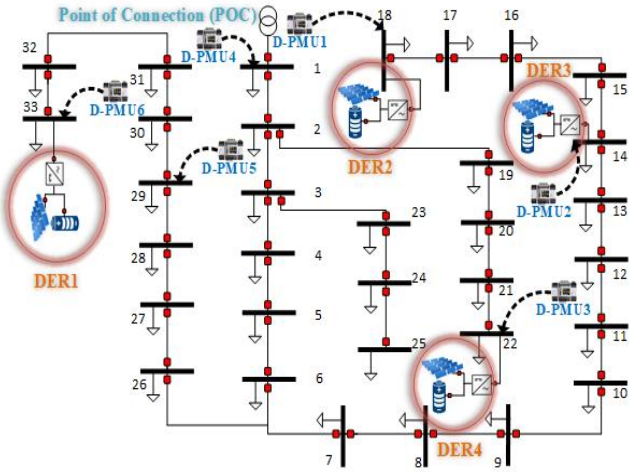


Fig. 4: The modified IEEE 33-node system with D-PMU and DERs

TABLE I: Stationary Operating Scenarios

Operating Scenario 1	All the renewables dispatching to the maximum capacity
Operating Scenario 2	All the four BESS discharging with 50% SOC
Operating Scenario 3	Zero renewable (loads being supplied by the grid)
Operating Scenario 4	BESS number 3 located at bus 22 is 50% charging

A series of load flow stationary scenarios as well as a comprehensive list of practical dynamic events have been modeled and simulated and all the time-stamped measurements are fed to the algorithms for efficacy investigations. The stationary use case development is to have the system operate under the edge scenarios of either all the power being supplied from the grid or the system maintains the load by its own in an islanded operation. Table-I shows a list of load flow operating scenarios, under which case several dynamic events have been simulated.

A wide range of dynamic use case scenarios have been modeled, including breaker operations, load and capacitor bank switching, generations step up and down, different types of short circuited bolted and shallow faults as well as islanding events. Corresponding D-PMU outputted measurements have been collected and pre-processed, further on timestamped measurements have been used for offline and online validations.

B. Simulation and Validation Analysis

For introducing noise and bad data to synthesis data from the OPAL-RT, the additive white Gaussian noise with the variance of $0.01 \times \mu_i$ added to each measurement i, where μ_i is the mean of that measurement. Non temporal-spatial anomaly bad data is acted uniformly to the measurements with rate 0.001 and change the value of the measurement

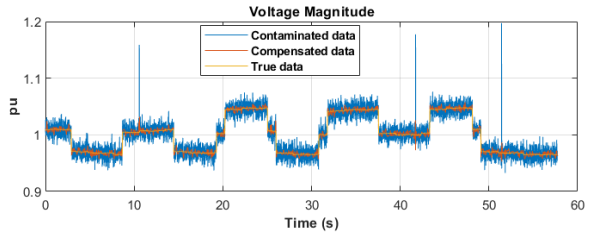


Fig. 5: The result of offline pre-processing for denoising and bad data compensation using MB-MLE and wavelet analysis on D-PMU 1, voltage magnitude measurement phase a

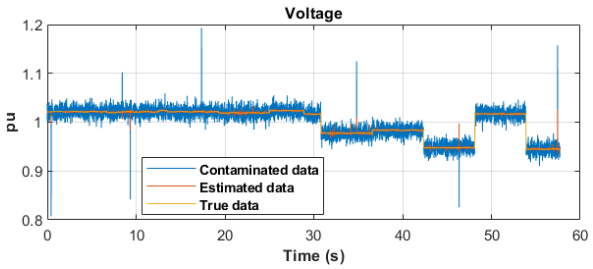


Fig. 6: The result of online denoising and bad data compensation using adaptive Kalman filter on D-PMU 2, voltage magnitude measurements, phase b

to value $\mu_i + \kappa_i$ where random variable κ_i is chosen from interval $[-\eta_i, -0.5\eta_i] \cup [0.5\eta_i, \eta_i]$ with uniform distribution where $\eta_i = max(10\sigma_i, 0.2\mu_i)$ and σ_i and μ_i are standard deviation and mean of measurement i respectively.

Fig. 5 shows the performance of suggested offline wavelet smoothing and MB-MLE BD anomaly detection and compensation technique on the contaminated data with noise and BD anomaly. As can be seen, all of the bad data are detected and eliminated from compensated data (at seconds 10.58, 41.73, 51.42), and also, the noise level has been reduced. The proposed offline tool increases the signal to noise ratio (SNR) 15.241 dB for D-PMU 1 voltage magnitude measurement phase a (Fig. 5). Table II shows the performance of each base anomaly detector and their MB-MLE assemble statistically in terms of the confusion matrix, recall, and precision. Recall is defined as $Recall = \frac{TP}{TP+FN}$ and precision is defined as $Precision = \frac{TP}{TP+FP}$. As can be seen, the MB-MLE has improved both Recall and Precision compare to base detectors.

TABLE II: Statistical comparison of different methods, Confusion matrix (True positive (TP), False positive (FP), False negative (FN), True Negative (TN)), Recall, and Precision

Method	TP	FP	FN	TN	Recall	Precision
HF	861	205	13	788713	0.9851	0.8077
QB	861	120	13	788798	0.9851	0.8777
DBSCAN	859	272	15	788646	0.9828	0.7595
MB-MLE	866	31	8	788887	0.9908	0.9654

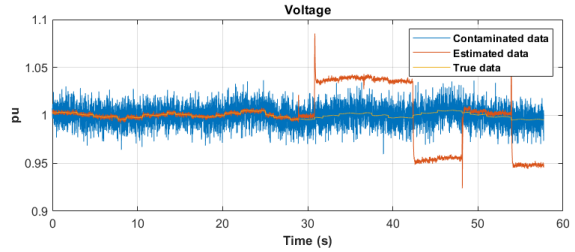
For online analysis, the signal is processed with offline tools and a linear dynamic is fitted to describe the linear dynamic. The linear dynamic is subsequently used to reduce the noise of the system and compensate for bad data (at seconds 0.41, 8.45, 9.31, 17.29, 14.09, 46.1, and 57.41). Fig. 6 shows that the introduced system can reduce the effect of noise and bad data significantly and SNR increases 9.78 dB for voltage measurement D-PMU 2 phase b (Fig. 6). Although with online KF, BD is not replaced through interpolation as it happens in the offline tool, it yet reduces the magnitude of BD anomaly significantly since it utilizes the KMA-based fitted model for prediction and filter uncorrelated spatial and temporal jumps in the measurement.

When the dynamic behavior has not been seen in fitting the linear model A, the process noise parameters of the system are adjusted adaptively to avoid relying on non-precise parameters for prediction and rely more on the noisy measurement. In Fig. 7 the performance of KF is shown when it is exposed to the dynamic events that have not seen before after 30 seconds. In 7a no compensation has been made for that and as can be seen, relying on an imprecise model leads to big bias. However, in 7b the deviation in the mean of estimated value and measurements are monitored and adjust the parameters of the KF. So after 30 seconds, we have a noisier estimation, however, we reduce the bias from the real value. That is because we trust less on the inexplicit model for prediction and rely more on the noisy measurement. 7c shows the ρ parameter where $Q = 0.1\rho I$, $R = I$ and I indicates the identity matrix with appropriate dimension.

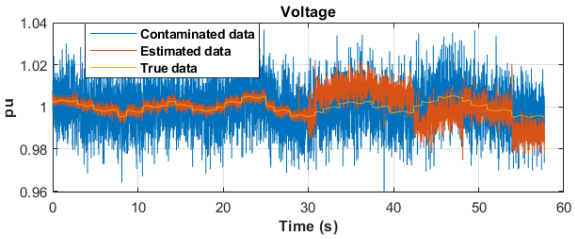
Although offline tool provides more precise denoising and BDD than online KMA-based KF filter, it comes with the cost of having more expensive computational time. For example, in our simulation using MATLAB software with Core i7 computer, for processing 114 data stream for 60 seconds with the sample rate of 120 data points per second, it takes 145.642 seconds for the offline tool to process the data. However, the same process is done in 2.592 seconds using the online tool. Therefore, based on the application, the appropriate tool can be selected for pre-processing of phasor measurements.

V. Conclusions and Future Work

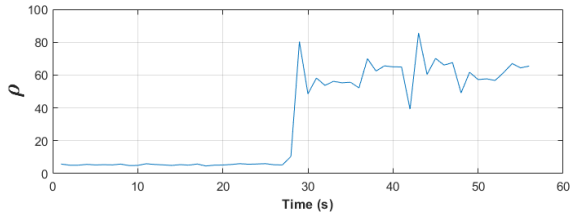
In this work, two new approaches are developed for offline (longer time window data processing) and online (short time window data processing) denoising and bad data detection in distribution Phasor Measurement Units (D-PMU). Multiple base detectors of different types including Hampel filter, Quartile detector and DBSCAN are developed and integrated using a margin-based maximum likelihood estimator (MB-MLE). High-frequency noises are processed and removed using a wavelet denoising method. Koopman Mode Analysis is used to fit a model to the underlying dynamics based on offline analysis and used for online denoising and bad data detection using Kalman Filter (KF) with adaptively adjusted parameters. Evaluation of performance using synthetic data generated by the modified IEEE 33 bus test system with multiple D-PMUs, simulated in real-time in the OPAL-RT confirms the practicality of the proposed scheme. For the next step, the validation can be further elevated using a large distribution network with a decentralized approach to divide the network into local zones to test the scalability of the proposed scheme. Also, online parameter adjustment of adaptive KF can be done



(a) without adaptive scheme



(b) with adaptive scheme



(c) parameter ρ

Fig. 7: Online denoising with D-PMU with Kalman Filter, voltage angle of D-PMU 3 phase a. after 30 second unforeseen dynamic act of the system

using deep autoencoder and dynamic features to enhance the fast response of the adjustment loop before deviation causes modification in the parameters.

REFERENCES

[1] S. Wang, L. Li, and P. Dehghanian, "Power grid online surveillance through PMU-embedded convolutional neural networks," in *2019 IEEE Industry Applications Society Annual Meeting*, 2019, pp. 1–8.

[2] J. Chen and A. Abur, "Placement of pmus to enable bad data detection in state estimation," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1608–1615, 2006.

[3] V. S. Bharath. Kurukuru, A. Haque, and M. A. Khan, "Fault classification for photovoltaic modules using thermography and image processing," in *2019 IEEE Industry Applications Society Annual Meeting*, 2019, pp. 1–6.

[4] A. Gholami, A. K. Srivastava, and S. Pandey, "Data-driven failure diagnosis in transmission protection system with multiple events and data anomalies," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 4, pp. 767–778, 2019.

[5] S. J. Hossain and S. Kamalasadan, "Combined deterministic-stochastic online subspace identification for power system mode estimation and oscillation classification," in *2019 IEEE Industry Applications Society Annual Meeting*, 2019, pp. 1–9.

[6] A. Gholami and A. K. Srivastava, "Comparative analysis of ml techniques for data-driven anomaly detection, classification and localization in distribution system," in *2021 North American Power Symposium (NAPS)*, 2021.

[7] M. Cramer, P. Goergens, and A. Schnettler, "Bad data detection and handling in distribution grid state estimation using artificial neural networks," in *2015 IEEE Eindhoven PowerTech*, 2015, pp. 1–6.

[8] Y. Peng, W. Qiao, L. Qu, and J. Wang, "Sensor fault detection and isolation for a wireless sensor network-based remote wind turbine condition monitoring system," in *2017 IEEE Industry Applications Society Annual Meeting*, 2017, pp. 1–7.

[9] S. Avdakovic, A. Nuhanovic, M. Kusljugic, and M. Music, "Wavelet transform applications in power system dynamics," *Electric Power Systems Research*, vol. 83, no. 1, pp. 237–245, 2012.

[10] P. Safayenikoo and I. Akturk, "Weight update skipping: Reducing training time for artificial neural networks," *CoRR*, vol. abs/2012.02792, 2020. [Online]. Available: https://arxiv.org/abs/2012.02792

[11] Y. An and D. Liu, "Multivariate gaussian-based false data detection against cyber-attacks," *IEEE Access*, vol. 7, pp. 119 804–119 812, 2019.

[12] F. A. Ghaleb, M. B. Kamat, M. Salleh, M. F. Rohani, and S. Abd Razak, "Two-stage motion artefact reduction algorithm for electrocardiogram using weighted adaptive noise cancelling and recursive hampel filter," *PloS one*, vol. 13, no. 11, p. e0207176, 2018.

[13] H. Liu, F. Hu, J. Su, X. Wei, and R. Qin, "Comparisons on kalman-filter-based dynamic state estimation algorithms of power systems," *IEEE Access*, vol. 8, pp. 51 035–51 043, 2020.

[14] I. Mezić, "Spectral properties of dynamical systems, model reduction and decompositions," *Nonlinear Dynamics*, vol. 41, no. 1-3, pp. 309–325, 2005.

[15] Y. Susuki, I. Mezic, F. Raak, and T. Hikihara, "Applied koopman operator theory for power systems technology," *Nonlinear Theory and Its Applications, IEICE*, vol. 7, no. 4, pp. 430–459, 2016.

[16] M. Netto, "Robust identification, estimation, and control of electric power systems using the koopman operator-theoretic framework," Ph.D. dissertation, Virginia Tech, 2019.

[17] R. Dubey, S. R. Samantaray, B. K. Panigrahi, and V. G. Venkoparao, "Koopman analysis based wide-area back-up protection and faulted line identification for series-compensated power network," *IEEE Systems Journal*, vol. 12, no. 3, pp. 2634–2644, 2016.

[18] S. P. Nandanoori, S. Kundu, S. Pal, K. Agarwal, and S. Choudhury, "Model-agnostic algorithm for real-time attack identification in power grid using koopman modes," *arXiv preprint arXiv:2007.11717*, 2020.

[19] A. Gholami, M. Mousavi, A. K. Srivastava, and A. Mehrizi-Sani, "Cyber-physical vulnerability and security analysis of power grid with hvdc line," in *2019 North American Power Symposium (NAPS)*, 2019, pp. 1–6.

[20] S. Pandey, S. Chanda, A. Srivastava, and R. Hovsapian, "Resiliency-driven proactive distribution system reconfiguration with synchrophasor data," *IEEE Transactions on Power Systems*, 2020.

[21] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of fluid mechanics*, vol. 656, pp. 5–28, 2010.

[22] M. R. Jovanović, P. J. Schmid, and J. W. Nichols, "Sparsity-promoting dynamic mode decomposition," *Physics of Fluids*, vol. 26, no. 2, p. 024103, 2014.

[23] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, "On dynamic mode decomposition: Theory and applications," *arXiv preprint arXiv:1312.0041*, 2013.