BAD DATA DETECTION AND IDENTIFICATION IN DISTRIBUTION POWER SYSTEMS BY MEANS OF PRINCIPAL COMPONENT ANALYSIS

Gerard Vancells Joaquim Meléndez Sergio Herraiz
University of Girona - Spain
sergio.herraiz@udg.edu

Juan Prieto Guillermo Bravo
INDRA - Spain
jprietov@indra.es

ABSTRACT

This paper focuses on the detection and identification of bad measurements, especially before power system state estimation. Some methods are described in the literature dealing with the most popular state estimation techniques based on weighted least squares (WLS). Here the Principal Components Analysis (PCA) algorithm is used to track the raw measurements, detect and identify the principal forms of gross and historic errors. The results show how single and multiple bad data are cleaned from raw measurements. The correct data can be introduced then in the state estimation algorithm allowing a more efficient estimation.

INTRODUCTION

Knowing with precision the state of an electrical power grid may be a challenging task. Currently distribution networks are operated almost without online measurements; this limits the efficiency of the exploitation. The deployment of vast quantities of instrumentation in future Smart Grids, if properly handle, will allow having a complete knowledge of the current state of the grid, improving drastically the efficiency of the operation. Nevertheless, only instrumentation is not enough to assure a reliable knowledge of the real-time state of the power system, because limited instrument precision, measurement errors and communication problems, may give an incomplete or inconsistent view of the grid state. For this reason, it is always necessary fit measurements to power systems equations in a process called 'state estimation', which searches for an 'average' or 'more probable' state [1].

It is well known that in this type of statistical problems (curve fitting problems) reliability of the result depends strongly on quality of raw data. Therefore, in order to get a consistent and reliable view of the grid state in future smart grids, raw data will have to be filtered and treated before feeding state estimation process and operation and management systems. This filtering, applied to future distribution smart grids that are massive in terms of number of elements and nodes, will foreseeable have to treat with vast amount of data in almost real-time.

The presence of bad data in raw measurements is usually due to the meter equipment or communication errors. Two principal types of bad data can be considered: single and multiple bad data. They are largely explained in [1].

The detection methods can work before or after the state estimation. The conventional method for state estimation is based on the weighted least squares WLS [1]. WLS is able to detect and identify single and multiple bad data by using the normalized residuals of the measurements, but it fails with conforming bad data. Post-estimation filters are widely studied in the literature, but not treated here.

Best known pre-estimation methods are based in autoregressive filters [4] or in Artificial Neural Networks [1], making a comparison between the measured value and the predicted value of measurements.

This paper presents a preestimation bad data detection based on the Principal Component Analysis (PCA), reducing the time and cost in comparison with ANN approaches, that could be very tedious and expensive [1].

PRINCIPAL COMPONENT ANALYSIS (PCA)

The Principal Component Analysis (PCA) is a method for reducing the number of main variables of a system, losing the least information. The PCA find relations between variables that describe the main data behavior, and builds a statistical model that represents the system. PCA also uses statistical indexes to detect and identify bad data or anomalous data [1]. Next, the procedure is explained.

Data matrix construction

The system data is organized in a matrix, Z:(*m* x *n*), with *n* measured variables and *m* observations, at every time step, of these variables. The matrix must be first scaled with zero mean and unit standard deviation.

Eigenvectors and Eigenvalues

The method is based on the covariance matrix decomposition into its eigenvectors.

$$\mathrm{cov}(Z) = S = \frac{Z^T Z}{m-1}_{S:(m\,\mathrm{x}\,n)} \quad (1)$$

The covariance matrix is diagonalized:

$$S\hat{P} = \hat{P}\lambda \quad (2)$$

where $\hat{P}$ is the eigenvectors matrix and λ is the eigenvalues matrix.

The eigenvectors are sorted from the highest eigenvalue to the lowest. The components with less importance are deleted, resulting a model with *a* principal components

retained and the *P* matrix "loadings" is obtained. Projecting the old variables through *P, T* matrix "scores" is calculated:

$$T = ZP \quad (3)$$

$$\hat{Z} = t_1 p^T{}_1 + t_2 p^T{}_2 + \ldots + t_a p^T{}_a \quad (4)$$

In this paper, a graphical method called scree test has been used to obtain the number of principal components, where the eigenvalues are plotted in a simple line plot and the place where the smooth decrease of eigenvalues appears to level off to the right of the plot is founded [7].

Statistic indexes Hotelling T^2 and Q

Hotelling T^2 index is defined as the distance from the principal components intersection to the measurement. T^2 defines an ellipse that contains the planes where the data is projected. The Hotelling T^2 is deduced as follows:

$$T^2 = \sum_{i=1}^{a} t_i \lambda_i^{-1} t_i^T \quad (6)$$

Q index is the variation of the data values outside the principal components included in the model. The Q or SPE index is deduced as:

$$Q = ee^T \quad \text{where} \quad e = Z - \hat{Z} \quad (7)$$

T^2 and Q limits

Thresholds off statistic indexes like T^2 can be calculated with a confidence α:

$$T_\alpha^2 = \frac{a(r-1)}{r(r-a)} F_{\alpha(a,r-a)} \quad (8)$$

Similarly, the way to get Q limit is:

$$Q_\alpha = \theta_1 \left[\frac{c_a h_o \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_o (h_o - 1)}{\theta_1^2} \right]^{1/h_0} \quad (9)$$

where

$$\theta_i = \sum_{j=a+1}^{m} \lambda_j^i \text{ for } i = 1,2,3 \quad \text{and} \quad h_0 = \frac{2\theta_1\theta_3}{2\theta_2^2} \quad (10)$$

Contribution analysis

When a limit is reached in an observation, it is possible to identify the original variables that are responsible of this anomalous situation. The method is able to know the contribution of each variable to the indices T^2 or Q.

Bad data replacement

If the method only detects and erases the bad data, the system could become unobservable, and the state estimation would not be performed. With PCA it is feasible to replace a bad data observed at time *t* with another observation got at time *t*-1. This is especially useful when the observations arrive close in time at the control center. Also, the variable could be reconstructed using the rest of measurements [8].

TEST SCENARIOS

Two scenarios have been considered to test the proposed preestimation technique: the IEEE 42-bus and a real HV/MV substation.

IEEE 42-bus system

The power system used for the simulations is the IEEE 42 nodes showed in Figure 1. This net is radial, and 5 buses have been considered for testing. At each bus, the measurements of voltages, active and reactive power have been obtained (V_a, V_b, V_c, P_a, Q_a, P_b, Q_b, P_c and Q_c) by using EMPT software. Sub indexes indicate the power system phase, a, b or c, being the total number of measured variables equal to 45. Varying the load consumption it is possible to get 18 different operation points. Noise can be introduced, simulating the meter behavior, into the observations. Then, if we create 10 extra observations from the 18 original observations, a total amount of 180 different observations are used to build the PCA model.

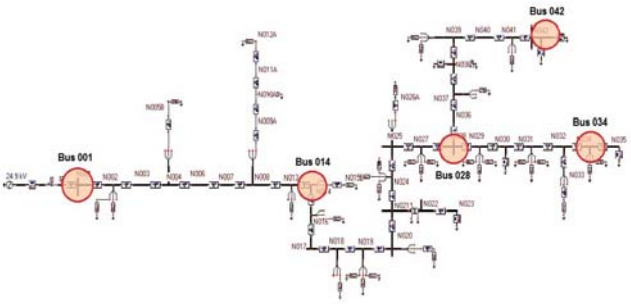

Figure 1. IEEE 42-bus system.

HV/MV substation

A real HV/MV substation has been considered as a second scenario, Figure 2.

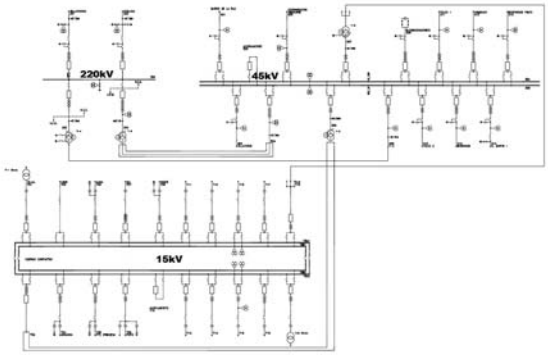

Figure 2. HV/MV substation.

The substation has three voltage levels, 220 kV, 45 kV and 15kV, and voltage, current, active and reactive power are measured at the different lines that feeds and are fed by the substation. In summary, 80 different variables are monitored.

The measurement matrix has originally 8760 measurements of 80 variables (24 measurements par day during a year). Before applying the PCA method, an exploratory analysis of the data was carried out to not consider erroneous data to build the PCA model (loss of data, data that not follow the basic electrical equations, etc.). As a result of this

prefiltering, a measurement matrix with 4126 observations and 69 variables is finally considered in the algorithm.

RESULTS

IEEE 42-bus system

Starting from the measurement matrix commented above, the PCA model is build retaining 44 principal components. Figure 3 shows the limit value T^2 (red line) considered a confidence $\alpha = 0.95$.

The PCA method works satisfactorily with both types of bad data, single and multiple. In the measurement matrix of 45 variables and 180 observations a bad data is introduced. The last row represents the last observation that has arrived to the center, so we suppose that past observations were been tested and were used to build the PCA model. For this reason, we only have to check the last observation.

Single bad data

The bad data introduced is the same measurement value plus ±20σ. The bad data is shown in Table 1.

Table 1. Single bad data construction

Variable	Real value	Sigma value σ	Error +20·σ	Bad data
V_a Bus 001 (V)	21737	58.9	1178	22915

In Figure 3, the blue line represents the values of the index T^2 for each observation. The last value corresponds to the observation with bad data included and it can be observed how index T^2 goes beyond the limit, detecting the bad observation.

In order to identify which variable has an error, the analysis of contributions is performed, which shows the weight of each variable in the value of T^2. Figure 4 shows the contributions for the bad observation, identifying the first one as a bad data, which is the voltage of the first node.

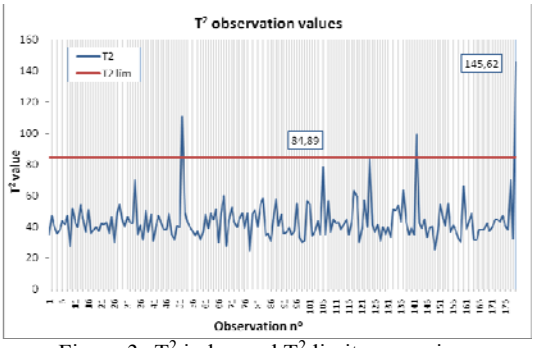

Figure 3. T^2 index and T^2 limit comparison

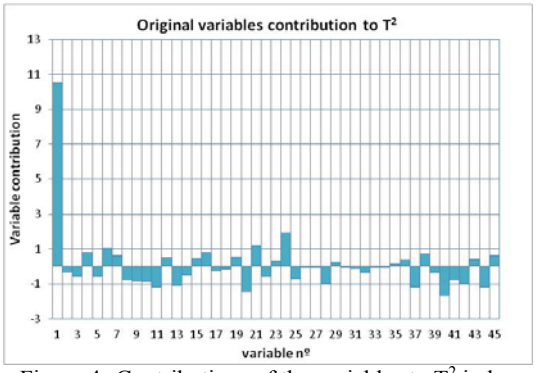

Figure 4. Contributions of the variables to T^2 index

Multiple bad data

A set of 5 bad data are introduced in this test, with the same characteristics that the single bad data.

Table 2. Multiple bad data construction.

Variable	Real value	Sigma value σ	Error +20·σ	Bad data
V_a Bus 001 (V)	21086	56.3	1126	22212
V_a Bus 014 (V)	18540	529	10580	29120
P_a Bus 028 (W)	55645	988	19760	75405
P_b Bus 028 (W)	99026	1491	29820	128846
V_b Bus 034 (V)	19773	950	19000	38773

First, the value of T^2 for the last observation (with errors) is compared with the value T^2 limit. As the limit is exceeded, the bad data is identified, Figure 5.

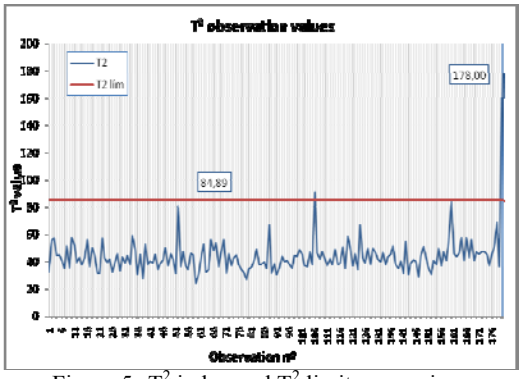

Figure 5. T^2 index and T^2 limit comparison

Next, the analysis of the contributions of each variable to T^2 could identify which variable has an error, Figure 6. In this example, the five bad data introduced in the observation are detected. However, the method should be iteratively applied, replacing the data with the biggest contribution for an estimated value (in this case, the last correct measurement). In this test, five iterations would be needed to identify the five bad data. Figure 7 shows the evolution of index T^2 in each iteration.

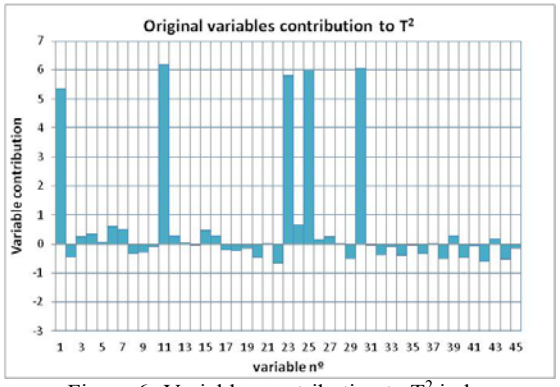

Figure 6. Variables contribution to T^2 index

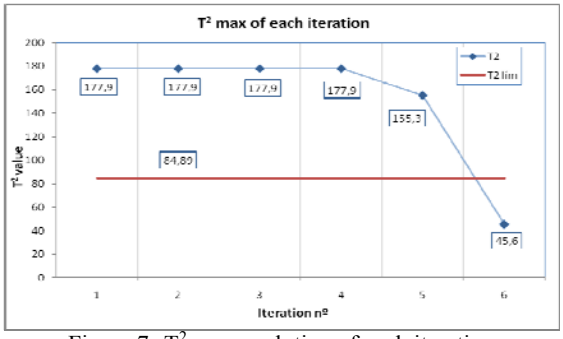

Figure 7. T^2 max evolution of each iteration

HV/MV substation

PCA model is built and, by using the scree test criteria [7], the number of the principal components to be retained in the model should be 5 or 12. Observations with bad data were generated starting from the prefiltered measurements, adding an error in each variable of each observation. The error was generated by using the standard deviation of each variable, σ, as is shown in Tables 3 and 4. In this case, only single bad data detection is presented.

Table 3 shows the performance of detection of the bad data generated when 5 principal components are retained in the model. It can be seen that for errors above 7·σ are successfully detected. Below this value, a percentage of bad data is not detected, for example, 58 % of the observations with errors equal to 6·σ. If the number of principal components is increased to 12, the algorithm is more sensitive to detect bad data. As it can be seen in Table 4, it can be detected errors equal to 4·σ (71% of cases tested).

Table 3. Single bad data detection performance (5 PC)

Error	True Positive	False Positive	True Negative	False Negative
5·σ	0%	0%	100%	100%
6·σ	42%	0%	100%	58%
7·σ	100%	0%	100%	0%
15·σ	100%	0%	100%	0%

Table 4. Single bad data detection performance (12 PC)

Error	True Positive	False Positive	True Negative	False Negative
3·σ	0%	0%	100%	100%
4·σ	71%	0%	100%	29%
5·σ	100%	0%	100%	0%
15·σ	100%	0%	100%	0%

CONCLUSIONS

PCA algorithm detects and identifies the principal forms of gross and historic errors before state estimation, allowing a faster and more reliable convergence. The PCA algorithm works satisfactorily with both types of bad data, single and multiple, and results show how single and multiple bad data are detected even when the deviation of the bad data is lower than ±4σ. The technique does not require information about the topology and parameters of the system, can work in real time, but the results show how the detection sensitivity depends obviously of the data that have been used to build the model.

ACKNOWLEDGEMENTS

This work has been supported by the research project "ENERGOS" (CEN-20091048) from the Program CENIT, CDTI, Ministry of Industry, Spain, and the research project "Monitorización Inteligente de la Calidad de la Energía Eléctrica" (DPI2009-07891) from the Ministry of Science and Innovation, Spain.

REFERENCES

[1] A. Gómez-Expósito et al., 2009, *Electric Energy Systems: Analysis and Operation*, CRC Press.

[2] A. Monticelli, F. Wu, M. Yen, 1986, "Multiple bad data identification for state estimation by combinatorial optimization", IEEE TPWRD, Vol. 1, 361-369.

[3] A. Monticelli, 2000, "Electric power system state estimation", Proceedings of the IEEE, Vol. 88.

[4] A. Abur, A. Keyhani, H. Bakhtiari, 1987, "Autoregressive filters for the identification and replacement of bad data in power systems state estimation", IEEE TPWRS, Vol. 2, 552-558.

[5] H. Salehfar, R. Zhao, 1995, "A neural network pre-estimation filter for bad data detection and identification in power system state estimation", Electric Power Systems Research, Vol. 34, 127-134.

[6] J.E. Jackson, 1991, *A User's Guide to Principal Components*, Wiley-Interscience.

[7] R.B. Cattell, 1966, "The scree test for the number of factors", Multiv. Behav. Res. Vol. 1, No. 2, 245-276.

[8] R. Dunia, S. Joe Qin, 1998, "Subspace approach to multidimensional fault identification and reconstruction", AIChE Journal, Vol. 44, 1813-1831.