Outliers’ Detection and Filling Algorithms for Smart Metering Centers

Ricardo Mion do Nascimento, Ana Paula Oening, Débora Cíntia Marcilio, Alexandre Rasi Aoki,
Eloy de Paula Rocha Júnior and Jefferson Marcelo Schiochet

Abstract—This paper summarizes diverse outliers’ detection and filling algorithms applied to metering centers of smart grids. The paper first shows typical problems in data measuring, especially considering smart metering infrastructure. The objective is to discuss and test detection and filling algorithms in several load time series of a real distribution utility. The impact of these algorithms is evaluated under a data mining approach, and results demonstrate the benefits that can be obtained in other applications like, load forecasting.

Index Terms—Outlier. Data Mining. Smart Grid. Smart Metering Centers.

I. INTRODUCTION

For any power system, the availability of information is essential for planning, operation and also for decision making. The lack of information or errors in measurement data can lead to failures in power supply and even put lives in danger.

The next generation of distribution systems as described in [1] has renewed interest in integrating all the system assets, including network equipment, consumers, and distributed generation sources under the same infrastructure for measurement, control and operation. A major challenge to achieve full functionality of smart grids concerns the development of a smart metering center. This difficulty includes defining the role and responsibility of that entity, for example, [2] presents a view of the control centers, monitoring and analysis for intelligent networks.

The biggest problem that involves the measurement centers concerns in the measurement data, which are redundant and reach the metering centers corrupted by failures in communication channels, lack of calibration of measuring equipment, equipment operating near the region of saturation or simply data missing, as presented in [3].

Another important aspect is about information security [4], because all the intelligence of the features provided in smart grids is based on the quality and accuracy of information, a critical flaw in the system or a cyber-attack to metering center [5], would result in the loss or corruption of information and this could make all the projected benefits of the smart grid fall down, generating huge losses for utilities and consumers and probably breaking laws imposed by regulators.

With the implementation of Smart Grid, the distribution utilities will demand a much larger amount of information for the operation and control of the grid [6] - [7]. This demand for greater quantities of electrical data can be solved by expansion of the supervised network with the installation of smart meters, yet the outliers in measurement data continue to occur and a system capable of verifying and correcting data contaminated by measurement problems is essential.

This paper presents several algorithms to detect and filling outliers applicable to metering centers, and is organized as follows: section 2 presents the main types of problems encountered in measurement data, section 3 shows some algorithms for detection and filling outliers, including an modified algorithm proposed by the authors, section 4 brings tests and results using load time series of a real Brazilian distribution utility, and section 5 draws the main conclusions and some future recommendations.

II. TYPICAL MEASUREMENT PROBLEMS

The most common problems encountered in measurement data is the absence data (nulls values and zeros), change in level and spikes (points more than M times the Standard Deviations away from the series mean) and generally are called outliers [8].

Sometimes the absence of data means, for example, a blackout, but in general is associated with measurement problems. Depending on the type of study or analysis, the correction of measurement data becomes essential, since basic statistics can be biased and thus lead to wrong conclusions.

The absence of data is usually characterized as an absence of values for a long time (Figure 1). Usually this failure is associated with problems in the communication system between the metering center and metering equipment or even on the measurement equipment. In the database, these records do not exist but for the purpose of study and other analysis, such as load forecasting, these data should be considered for the construction of the series.
Fig. 1. Missing data (null values or zeros).

Fig. 2. Change of level.

Fig. 3. Spikes.

Many techniques for replacing or filling can be employed and the application of any of these techniques depends on tests to verify the adherence of the data calculated in the original series.

The level change is identified as an abrupt change in the pattern or in the average value of the series and it tends to remain for a certain period (Figure 2). Often, the level change occurs due to changes in network topology (switchings or contingency of other equipment’s system) which turns out to redistribute power flows in network operation. Depending on the purpose of the study or analysis, these data don’t need correction, since failures are not due to measurement problems.

A spike (Figure 3) is an observation that differs from the expected value of other observations for the period. One of the probable causes of the occurrence of spikes is momentary instability of the sampled signal during transduction. This type of outlier should be corrected regardless of the study or analysis to be performed on the data.

In addition, there are other causes of outliers in measurement data like, for example, current and voltage transformers saturation, measuring or transduction equipment damaged, parameterization or improper installation, registration problems in the database, etc.

III. TECHNIQUES FOR DETECTION AND FILLING OF DATA

Depending on their nature, the outliers may not cause any major effect but also can cause considerable impact during the data analysis. Thus, the importance of detecting outliers lies in understanding the series under analysis: the detection of an outlier can be an evidence of the occurrence of some external factor affecting the series.

There are numerous methods in the literature used for outliers’ detection based on conventional techniques such as Likelihood Ratio Test, Score Test (or Lagrange Multiplier Test), Likelihood Displacement and also techniques like neural networks.

The great problem, however, lies in the fact that some techniques have better performance than others depending on the nature of the series or type of outliers.

This study evaluated and tested methods based on conventional techniques for the detection of outliers, such as:

-- Extreme Studentized Deviate (ESD);
-- Generalized Extreme Studentized Deviate (GESD);
-- Z-Score;
-- Modified Z-Score;
-- Test Box Plot;
-- Thompson;
-- Adjusted Boxplot;
-- Exponential smoothing (ExpSM).

In the series analyzed, the most noticeable problem is the presence of null values (translated by the absence of measurement or information) which is also the largest contributor to the poor performance of the methods tested. Therefore some tests involved both the automatic filling of periods characterized by the absence of data and also the data standardization using Box-Cox transformation.

In general, there is a difficulty in the stage of setting the method’s parameters. It varies depending on the pattern of the analyzed series and the degree of degradation of the same, i.e., certain setting can work very well for outliers’ detection for a series in a given period and yet this setting may not be sufficient for detecting outliers in other series, or even identify wrongly correct observations as outliers. Table I shows the performance of the methods for each analyzed series.

Disregarding the difficulty of adjusting the parameters, the method modified Z-Score showed the best performance. However, it was found that some aspects could be further discussed to improve the performance of the method, such as:

-- Correction of null values: the presence of zeros influences the less robust statistics as mean (x̄) and standard deviation (s);
TABLE I

PERFORMANCE OF OUTLIERS’ DETECTION AND FILLING METHODS.

<table>
<thead>
<tr>
<th>Series 1</th>
<th>Series 2</th>
<th>Series 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESD</td>
<td>good</td>
<td>bad</td>
</tr>
<tr>
<td>GESD</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Z-Score</td>
<td>good</td>
<td>bad</td>
</tr>
<tr>
<td>Modified Z-Score</td>
<td>good</td>
<td>good</td>
</tr>
<tr>
<td>Test Box Plot</td>
<td>bad</td>
<td>bad</td>
</tr>
<tr>
<td>Thompson</td>
<td>good</td>
<td>bad</td>
</tr>
<tr>
<td>Adjusted Box Plot</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>ExpSM</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

-- Verification of normal distribution and stationary of the series: many methods require that the series analyzed is stationary and has a normal distribution. It is known that null values and outliers contribute to the distortion of the distribution of the series;

-- Period of analysis: the size of the period under analysis may be insufficient for the performance of the method. Thus, the use of a longer period, known and free of outliers, can help to increase method’s performance;

-- Series breakdown by type of day (clustering): grouping load curves according to the type of day of week (weekday, Saturday and Sunday/holiday) can reduce the variability of the data subsets and thereby improve methods’ performance.

A. Modified Z-Score

The problem of modified Z-Score method is that \( \bar{x} \) and \( s \) can be greatly affected by outliers. An alternative is to replace them with more robust estimators, i.e., replace \( \bar{x} \) by sampled mean (\( \bar{x} \)) and \( s \) by MAD (median absolute deviation).

\[
\text{MAD} = \text{median}(|x_i - \bar{x}|)
\]

Thus, the method modified Z-Score is defined as:

\[
M_i = \frac{0.6745}{\text{MAD}} (x_i - \bar{x})
\]

The observations are identified as outliers when \( |M_i| > D \), in the simulations was considered \( D = 3 \), as the value that has the best performance to deal with the types of errors and the nature of the analyzed series.

B. Filling Algorithms

Outliers’ Filling algorithms involved different average calculations considering different combination of prior periods in order to get a strategy that was able of producing consistent observations from the history of the series. Considering a load series with hourly observation, the following algorithms were:

diff(w_pre+h_pre): difference between two observations corresponding to the same period of previous week plus the value of last observation.

\[
T_h = T_{h-1} + (T_{h-168} - T_{h-167})
\]

where \( T_h \) is the estimated observation, \( T_{h-1} \) is the last observation, \( T_{h-168} \) and \( T_{h-167} \) are observations referent to the same period of previous week, respectively.

\[
\text{M}(w_{pre}+h_{pre}): \text{average of last observation and previous week observation.}
\]

\[
T_h = (T_{h-1} + T_{h-168})/2
\]

\[
\text{WM}(d_{pre}+h_{pre}): \text{weighted mean of last observation and the observation of previous day.}
\]

\[
T_h = (p_1 \cdot T_{h-24} + p_2 \cdot T_{h-48})/(p_1 + p_2)
\]

where \( p_1 \) and \( p_2 \) are the weights.

\[
\text{M}(2d_{pre}): \text{average of two previous days.}
\]

\[
T_h = (T_{h-24} + T_{h-48})/2
\]

\[
\text{WM}(2d_{pre}): \text{weighted average of two previous days.}
\]

\[
T_h = (p_1 \cdot T_{h-24} + p_2 \cdot T_{h-48})/(p_1 + p_2)
\]

\[
\text{M}(3d_{pre}): \text{average of three previous days.}
\]

\[
T_h = (T_{h-24} + T_{h-48} + T_{h-72})/3
\]

\[
\text{WM}(3d_{pre}): \text{weighted average of three previous days.}
\]

\[
T_h = (p_1 \cdot T_{h-24} + p_2 \cdot T_{h-48} + p_3 \cdot T_{h-72})/(p_1 + p_2 + p_3)
\]

where \( p_1 \), \( p_2 \) and \( p_3 \) are the weights.

\[
\text{M}(2d_{pre}+h_{pre}): \text{average of two previous days and last hour.}
\]

\[
T_h = (T_{h-24} + T_{h-48} + T_{h-1})/3
\]

\[
\text{WM}(2d_{pre}+h_{pre}): \text{weighted average of two previous days and last hour.}
\]

\[
T_h = (p_1 \cdot T_{h-24} + p_2 \cdot T_{h-48} + p_3 \cdot T_{h-1})/(p_1 + p_2 + p_3)
\]

\[
\text{M}(3d_{pre}+h_{pre}): \text{average of three previous days and last hour.}
\]

\[
T_h = (T_{h-24} + T_{h-48} + T_{h-72} + T_{h-1})/4
\]

\[
\text{WM}(3d_{pre}+h_{pre}): \text{weighted average of three previous days and last hour.}
\]

\[
T_h = (p_1 \cdot T_{h-24} + p_2 \cdot T_{h-48} + p_3 \cdot T_{h-72} + p_4 \cdot T_h)/(p_1 + p_2 + p_3 + p_4)
\]
where $p_1$, $p_2$, $p_3$ and $p_4$ are the weights.

The study for algorithm’s choose able of generating observations of a series from its history demanded the selection of a period for the application of the algorithm and other period for validation.

We selected periods of several series with different characteristics to verify the algorithm that best reproduce the nature of the series from its history. To illustrate this process of choice of the algorithm, it’s shown two series (a) and (b). After application of the algorithm in the corresponding period and validation, we calculated the residues between the original value of the series and the value obtained for each algorithm.

Table II shows the quadratic summation of these residues for series (a) and (b). It’s possible to conclude that the best algorithms are $\text{dif}(\text{w}_\text{pre}) + h_{\text{pre}}$, $\text{M}(2d_{\text{pre}})$ e $\text{WM}(3d_{\text{pre}})$.

The Figures 4 and 5 show series (a) and (b) where the validation’s load curves (MW) are compared to obtained load curves for each tested algorithm.

<table>
<thead>
<tr>
<th></th>
<th>series (a)</th>
<th>series (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{dif}(\text{w}<em>\text{pre}) + h</em>{\text{pre}}$</td>
<td>1.69</td>
<td>18.91</td>
</tr>
<tr>
<td>$\text{M}(2d_{\text{pre}})$</td>
<td>1.49</td>
<td>45.31</td>
</tr>
<tr>
<td>$\text{WM}(2d_{\text{pre}})$</td>
<td>1.59</td>
<td>46.70</td>
</tr>
<tr>
<td>$\text{M}(2d_{\text{pre}} + h_{\text{pre}})$</td>
<td>62.25</td>
<td>65.04</td>
</tr>
<tr>
<td>$\text{WM}(2d_{\text{pre}} + h_{\text{pre}})$</td>
<td>133.81</td>
<td>103.66</td>
</tr>
<tr>
<td>$\text{M}(3d_{\text{pre}})$</td>
<td>1.32</td>
<td>33.82</td>
</tr>
<tr>
<td>$\text{WM}(3d_{\text{pre}})$</td>
<td>1.39</td>
<td>37.65</td>
</tr>
<tr>
<td>$\text{M}(3d_{\text{pre}} + h_{\text{pre}})$</td>
<td>23.68</td>
<td>46.99</td>
</tr>
<tr>
<td>$\text{WM}(3d_{\text{pre}} + h_{\text{pre}})$</td>
<td>74.38</td>
<td>67.37</td>
</tr>
<tr>
<td>$\text{M}(\text{w}<em>{\text{pre}} + h</em>{\text{pre}})$</td>
<td>41.55</td>
<td>48.01</td>
</tr>
<tr>
<td>$\text{M}(d_{\text{pre}} + h_{\text{pre}})$</td>
<td>235.46</td>
<td>188.18</td>
</tr>
</tbody>
</table>
C. Proposed Method

The proposed method involves a step of pre-detection of outliers using the method modified Z-Score and also a post-detection step, which can be characterized by:

Pre-Detection

-- Defining the period of historical data: corrected and free of outliers data set for comparison with data for the period under analysis. For testing the historical period used was two times greater than the period of data analyzed, i.e., 2 months.
-- Clustering of data according to the type of day of week (weekday, Saturday and Sunday/holiday);
-- Zeros and null values: days with more than 50% of data with zeros and nulls values are excluded from analysis, otherwise the outliers are replaced by values calculated by some of the algorithms presented in section B.

Post-Detection

-- The observations identified as outliers by the method modified Z-Score are replaced by values calculated using one of the algorithms in section B;
-- Compare the calculated values with the values of the original series, the difference between them is greater than a threshold to be defined, this observation is classified as an outlier and it will be replaced by a calculated value, otherwise the original value of the series is maintained.

IV. TESTS AND RESULTS

The tests were made using power flow series for substation equipment with predominantly urban and residential characteristics. Each of these figures is the result of applying the methods on the same series shown in Figures 1, 2 and 3. These series were first analyzed by modified Z-Score method and then the modified Z-Score with the steps of pre/post-detection. In the figures below, the dashed line represents the sequence of new values for observations identified as outliers.

Figures 6 and 7 refer to a series of power flow obtained by measuring a transformer in November 2010. In Figure 6 we applied the modified Z-Score method and Figure 7 shows the application of the modified Z-Score with pre/post-detection step. It's easy to note that the modified Z-Score method with pre/post-detection step identified a larger number of outliers, including the observations occurred in the day November 15th (Monday and national holiday). One of the observations of November 21st (Sunday) was also classified as an outlier, but this occurred because the load that Sunday was atypical compared to the other Sunday.

In Figures 8 and 9, the analyzed series refers to the active power flow from a feeder with characteristic urban, residential and industrial, in June 2010. In Figure 8, the modified Z-Score method only identified as outliers lowest values occurred during the period from June 11th to 18th. The modified Z-Score method with pre/post detection step, not only identified these observations as outliers, but also identified the discrepancies occurred on days June 4th, 27th and 29th.

Figures 10 and 11 refer to a series of power flow from a transformer in April 2010. In Figure 10 it’s shown that the spikes occurred on days April 15th and 18th were detected, but some correct observations were also wrongly classified as outliers. In the Figure 11, the modified Z-Score method with pre/post-detection steps identified only spikes cited and also part of the observations of April 2nd, because of a untypical load on this day.

V. CONCLUSION

The smart metering centers should be prepared for a huge amount of measurement data from the smart grids of all types, including consumer smart metering, power quality measurements, information about the state of network equipment, electrical quantities in several network measuring points, and measurements of distributed generation sources, among others.

Most of these measurements will contain errors of various types, like failures in communication channels or in measuring equipment and data corruption that are the most common. Therefore, to allow an effective operation and control of smart grids is necessary to apply data mining techniques to improve the quality of measurement data used by multiple applications in their decision-making processes.

This paper presented techniques for detecting outliers and filling applied to several time series of a real load Brazilian distribution utility, demonstrating its need, applicability and effectiveness.

These algorithms application seek to improve the performance of applications that use these data as inputs, such as load forecasters, power system analyzers, among others.

Future work includes algorithm development for self-tuning of the algorithms presented in order to improve the performance of the automatic application of the techniques.

REFERENCES

Fig. 6. Method Modified Z-Score - Missing data (null values or zeros).

Fig. 7. Method Modified Z-Score + pre/post-detection steps - Missing data (null values or zeros).

Fig. 8. Method Modified Z-Score - Change of level.

Fig. 9. Method Modified Z-Score + pre/post-detection steps - Change of level.

Fig. 10. Method Modified Z-Score - Spikes.

Fig. 11. Method Modified Z-Score + pre/post-detection steps - Spikes.