DOI: 10.48011/asba.v2i1.968

Data Analysis and Preprocessing Method of Medium Voltage Distribution Network Feeders

Bacalhau, J. M. R. * **Fardin, Jussara** **

* *Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal do Espírito Santo, ES,(e-mail: joao.bacalhau@aluno.ufes.br)*
** *Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal do Espírito Santo, ES,(e-mail: jussara.fardin@ufes.edu.br)*

Abstract: The investment in the energy sector aims to ensure a continuous, reliable, and quality supply of electrical energy imposed by the electricity regulatory agency with maximum economic-financial balance. This paper discusses the challenges of processing data from medium voltage distribution feeders to use on the distribution network planning. The analysis of missing data and outliers is made on the three-phase voltage, current, and power factor of 459 time series of real feeders. Furthermore, it is proposed a method of preprocessing, and missing data imputation using the unbalanced characteristic between phases, interpolation, and the normalized scaled standard weekday curve. The results show that most missing data are three-phase, however, with a significant amount of single and dual-phase loss that can be filled by the proportion between phases. Hence, the challenge is to fill multiple weeks of missing three-phase data, and for that, it is proposed the use of the standard curve for each day of the week. The method proposed is a promising alternative for data imputation in medium-voltage feeders. The technique is tested using real feeder data degraded by its missing data probability function, and compared with the Naïve approach.

Keywords: Network expansion; Distribution system planning; Data imputation; Feeder; Data analysis; Missing value; Incomplete data; Imputation; Time series data.

1. INTRODUCTION

1.1 Power distribution network planning

DISTRIBUTION networks are the last mile on the delivery of energy from the generators to the end-users. Typically, following a radial construction, different feeders come out from the substation running across many different areas. By far, this part of the system has the highest complexity level because of its extension, the number of equipment, variability of load characteristics, and the number of possible reconfigurations. Since its an essential part of the process, the amount of investments in the distribution network is very high, hence, it demands careful planning (Gonen, 2007; Muñoz-Delgado et al., 2018).

Regarding the utility company, the power distribution planning is of extreme relevance as it is responsible for increasing the capability of the system maintaining a continuous, reliable, and quality supply of electrical energy (Vargas, 2015). The distribution planning relies on the quality of the information and availability to make decisions on the sector. Therefore, the lack of reliable data directly impacts the strategic objectives of electricity companies and the efficiency of the investments of medium and long term.

⋆ This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

1.2 Data analysis and missing data imputation

The analysis of historical information is a powerful tool to discover trends and patterns in businesses (Han et al., 2012). The process of acquisition and storage of data from the electric distribution system has a chain of actors since the physical measure until the storage in the Distribution System Operator (DSO). Furthermore, each step of the process is subject to interference and, consequently, loss and alteration of the information. The major players in the process are the failure of equipment that alters measurements and weather conditions that prevent the transfer of information. These facts, together with the stop of equipment for preventive maintenance and load maneuvers (specifically in the context of system planning), cause outliers and vacancies to appear on the dataset.

Since inappropriate treatment of missing values may cause incorrect results in data mining, the problem of missing value imputation has become a focus in the analysis of incomplete data in opposition to sample deletion. Several imputation methods have been proposed, such as Imputation with constant, Mean Imputation, Hot Deck Imputation, Auto-Regression Models, Linear interpolation, Random Imputation based on statistical distributions, among others. Furthermore, k nearest neighbor, neural networks, support vector machines, and auto-encoders are among some of the new strategies (Peppanen et al., 2016; Jadhav et al., 2019; Saunders et al., 2006).

In this work, an imputation method is proposed based on the correlation of the time series studied and the weekdays. The method uses the normalized standard day of the week linearly scaled by the minimum and maximum values of samples on its time series.

1.3 Dataset

This research utilizes a dataset that consists of nine time series for each one of the 459 medium voltage (MV) feeders of a utility company in Brazil. The nine time series are the three-phase voltage (V), current (I), and power factor (pf) collected from January the 1st to December the 31st of 2019 sampled at the relays on the distribution substation. Similarly to other utilities, the portfolio of relays is vast and is composed of different manufactures and technologies. The main implication for the study is that some relays will sample at 5 min periods were others at 1 min. Besides the primary dataset other two secondary datasets were used as support. The first one has the topological data such as the feeder commissioning date, cable gauge at substation output, and the nominal voltage. The second one has the log of the load transfer between feeders. The last one has the start and the end timestamps of the load transfers made on the distribution network and the feeders that were involved.

In this article, all the discussion is focused on primary information (V, I, pf), given that theoretical relationships can calculate secondary information. For instance, the active/reactive power and the energy can be calculated if, for every timestamp, the voltage, current, and power factor information exists. Naturally, this assumption requires that there is no missing data on the primary measurements dataset. Thus, the importance of the proposed work.

1.4 Context and contribution

This paper, the challenge of using data from medium voltage distribution feeders as input for power distribution planning is discussed. The analysis of missing data and outliers is made on the three-phase data of voltage, current, and power factor of 459 time series of real feeders. Furthermore, it is proposed a method of preprocessing with outlier removal, and missing data imputation using the unbalanced characteristic between phases, interpolation, and the scaled normalized standard curve for each day of the week.

The method proposed is tested using real data degraded by its missing data probability function. The preprocessing and imputation method proposed is discussed and the last compared with the Naive approach. It is important to cite that all the work that is presented was implemented using python 3.7 and libraries such as, but not limited to, pandas, NumPy, and matplotlib.

2. METHODS

2.1 Time series synchronization

The time series synchronization is the first step in processing the dataset. The synchronization is vital since the alignment between phases of the same quantity, between quantities of the same feeder and between feeders, provides many advantages as described. The first one being the ability to combine all nine time series of each feeder and calculate the secondary quantities time series ($P_{active/reactive}, E_{active/reactive}$). Furthermore, the synchronization between feeders provides the capability to analyze the iteration between them, for instance, in load transfers for scheduled maintenance and to estimate quantities of the substation transformers by the sum of all feeders related. Finally, in the dataset, there are two different time sampling periods, being the most prolonged and predominant 5 min, all the feeders that are sampled at 1 min period were downsampled to 5 min. Thus, the process also reduces the size of the dataset.

	Sample	Timestamp	ϕ_a	ϕ_b	ϕ_v
1sec Shift ↑	1	01/01/19 00:00:01	14,29	↑	14,29
	2	~~01/01/19 00:01:49~~		14,10	
	3	~~01/01/19 00:02:01~~			~~14,29~~
	4	~~01/01/19 00:08:01~~	~~14,32~~	~~14,13~~	~~14,34~~
	5	01/01/19 00:09:48	14,32	~~14,12~~	14,32
	6	01/01/19 00:10:01	↓	14,31	↓

⇩

Sample	Timestamp	ϕ_a	ϕ_b	ϕ_v
1	01/01/19 00:00:00	14,29	14,10	14,29
-	01/01/19 00:05:00			
6	01/01/19 00:10:00	14,32	14,31	14,32

Figure 1. Raw data synchronization and downsampling for three-phase voltage.

Figure 1 shows a three-phase time series slice of the voltage from a feeder and the process of synchronizing and downsampling. To synchronize the samples of different feeders all the time series are shifted to start at 01/01/2019 00:00. In the example, all of the three phases were shifted by one second, and samples three and four were discarded. Furthermore, phases A and V from sample 5 were used to fill the missing data from sample 6, and phase B from sample 2 was used to fill sample one. It is assumed that for an interval of up to 5 min, the variance would be negligible, and the use of timestamps $i + 1$ and $i - 1$ to fill gaps in timestamp i would not compromise the analysis. On the bottom of Fig. 1, the result of the process is shown. Although there were six samples in the raw data, the result has only two complete three-phase samples, and the timestamp of 01/01/19 00:05:00 is missing. In further steps, the algorithm proposed will insert the resultant missing samples. Nevertheless, the three main steps in this part of the process are: synchronizing the start of the time series, utilize samples $i+1$ and $i-1$ to fill missing data in sample i, as long as they are less than half of the period distant, and downsampling. In this part of the process, the start and the end of the period of study are defined. This outline is essential as each time series is synchronized with the starting point, and all samples collected after the end are discarded.

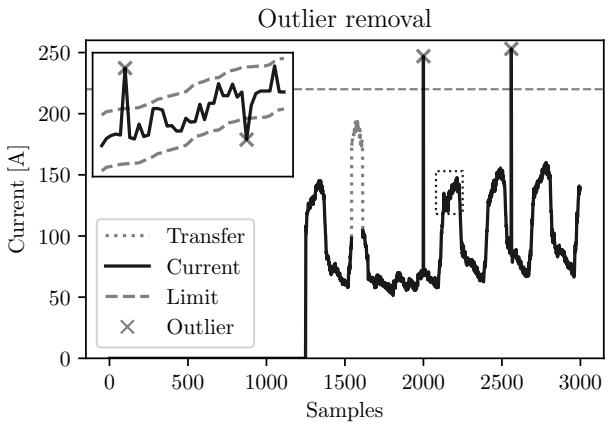

Figure 2. Example of outlier removal for current.

Another relevant aspect is to determine the beginning of the operation of each medium voltage feeder through the topological dataset. The start is important as the distribution network changes with time, and new substations and or feeders can be commissioned in the middle of the period of study. Therefore, since the start of the period of study until the commissioning of the feeder, all time series samples should be set to zero. The information about the beginning of the operation ensures that an imputation algorithm would not input data in a period where the feeder did not exist. For instance, the feeder shown in Fig. 2, commissioned in sample 1250, had all the samples before that timestamp set to zero.

2.2 Outlier removal

The outlier removal, in the context of this work, can be split into three parts, as shown in Fig. 2. The first one is to remove data that was sampled during the load transfer between medium voltage feeders on the distribution network. These situations are considered anomalies as they do not represent the system's regular operation. The information from the dataset of load transfers is used and, for each period where a feeder received or gave way load, the samples of current and power factor are discarded. In Fig. 2 around sample 1500, the feeder receives load, and therefore this period is marked as not valid and removed. It is essential to mention that the samples from the voltage time series were not discarded. Statistically, the load transfer does not change the characteristics of the substation voltage bar, as shown in section 3.1. The second part refers to the physical and theoretical constraints of the system. For example, power factor samples that are not between zero and one or current sampled by the relay that is greater than the capacity of the cables at the substations output. As stated previously, the physical information of the feeders was obtained on the utility company's topological database.

Furthermore, voltage samples that are greater than 1.1 pu or lower than 0.9 pu are unrealistic in the regular operation of the system should be removed as outliers. For the case shown in Fig. 2, the physical capacity of the cable is 220A, therefore, two samples marked were removed as outliers.

By last, a statistical method was used for removing the remaining outliers. In LEYS et al. (2019), the authors state that it is common practice the use of plus and minus the standard deviation ($\pm\,\sigma$) around the mean (μ), however, this measurement is particularly sensitive to outliers. In this work, a variant of the method proposed by LEYS et al. (2019) was implemented. The limit was set by the median absolute deviation ($\pm\,MAD_i$) around the moving median (M_i) where i denotes the number of samples of the moving window. Typically, an MV feeder has a seasonality where in the summer load is higher than in the winter or vice-versa. Hence, it is vital to use the moving median instead of the median of all the time series. The top left corner of Fig. 2 shows the superior ($Max_{threshold}$) and inferior ($Min_{threshold}$) limits defined by (1).

$$\begin{aligned} Max_{threshold} &= M_i + N * MAD_i \\ Min_{threshold} &= M_i - N * MAD_i \end{aligned} \tag{1}$$

The length i of the window and the number of median absolute deviations denoted N were defined empirically, for each one of quantities analyzed (V, I, pf). In the example, two samples were marked as outliers as they were not in between acceptance limits.

2.3 Load transfer and bus voltage

The load transfer between medium voltage feeders, as stated by Wen-Chih Yang (2011), is an essential part of ensuring the reliability of the power distribution network. However, for planning the expansion of capacity for the system, all data collected during the temporary load transfers must be discarded. The effect of the load transfer for the current and power factor is very prominent. However, in the bus voltage of the substation, this is not true. Hence, the following procedure was conducted to verify that the load transfer between feeders on the distribution network did not change the bus voltage characteristics. For each one of the 115 MV buses in the dataset, the three-phase voltage average during the load transfer of any related feeder was compared with the average during normal operation. This was done using a dependent sample t-test with 5% α (Shier, 2004), the results are shown in section 3.1. Therefore, if there is no statistical difference for the bus voltage in the two cases mentioned, it is not required to remove the periods of load transfer from the voltage time series of MV feeders.

2.4 Imputation method proposed

The imputation method proposed is shown in algorithm 1. The process has three main parts: initial processing and interpolation, data filling based on the ratio between phases, and data filling based on the normalized scaled standard day of the week curve (NSSC). The initial part handles the data synchronization (section 2.1), outlier removal (section 2.2), and the first linear interpolation. The first linear interpolation, done individually for each quantity and phase, is limited by $N_{samples}$ in length. Empirically, for the dataset studied, it was assumed $N_{samples} = 18$ (1.5 hours) as for this number of samples, the characteristics of the voltage, current, and power factor do not change dramatically. As shown in section 3.2, this will make for the most of the data that is missing. However, the interpolation will not solve the most problematic case, which is when the number of consecutive missing values is large (days, weeks, and months).

Algorithm 1: Preprocessing and imputation method

Input: MV feeder dataset, Topological dataset, Load Transfer dataset and period of study start/end

for *each feeder* **do**
- Syncronize time series
- **for** V, I *and* pf **do**
  - Remove outliers
  - Apply linear interpolation ($N_{samples} = 18$)
  - **for** *each missing sample* i **do**
    - **if** $X^i_{\phi_a} = null$ *and* $X^i_{\phi_b} \wedge X_{\phi^i_v} \neq null$ **then**
      - Apply (2)
    - **if** $(X^i_{\phi_b} \vee X^i_{\phi_v}) \neq null$ **then**
      - Apply (3)
  - **for** *each phase* (ϕ) **do**
    - **if** *Every wd has at least 3 vd* **then**
      - Calculate the NSSC using (4)
    - **else**
      - Find equivalent feeder in dataset with at least 3 vd for each wd
      - Calculate the NSSC based of equivalent feeder using (4)
    - **for** *each day* (d) **do**
      - **if** *No missing samples* **then**
        - Calculate Min/Max vd values
        - Add Max of vd to γ vector
        - Add Min of vd to ζ vector
    - Coompute the moving average of two samples of γ and ζ vectors
    - **for** *each day* (d) **do**
      - **if** *Missing samples* $>= 50\%$ **then**
        - **if** *Between vd* **then**
          - Subst. day with (5)
        - **else**
          - Subst. day with (6)
      - **else**
        - **for** *each period* (p) *of day* (d) **do**
          - **if** *Missing samples* $>= 50\%$ **then**
            - **if** *Between vd* **then**
              - Subst. part of day with (5)
            - **else**
              - Subst. part of day with (6)
  - Final linear interpolation ($N_{samples} = \infty$)

return

After the first interpolation, the second stage uses the correlation between phases (ϕ_a, ϕ_b, ϕ_v) of the same quantity (V, I, pf) to infer a missing sample value based on adjacent samples. Adjacent samples are those of the same timestamp i but from different phases that the one which is missing. In this step, different periods T of analysis are considered (dawn, morning, afternoon, night, month, and year). Where T^i denotes the part of the day (dawn, morning, afternoon, and night), month or year in which the sample $X^i_{\phi_a}$ is contained. Therefore, if the period T^i of all

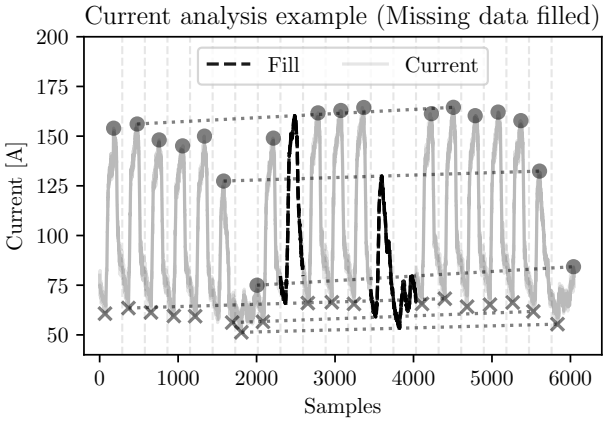

Figure 3. Example of using the normalized scaled standard weekday curve method, (4), (5), and (6), to fill missing days. The circles indicates the maximum values (γ) whereas the crosses indicates the minimum values (ζ) for each valid day (vd).

the three-phases (ϕ_a, ϕ_b, ϕ_v) has more than 50% of data, the ratio between the phases is calculated and the missing sample is filled, refer to (2) and (3). If there is not enough data in part of the day (p), the process is repeated for the month (m) and then for the year (y) in which the missing sample is contained. The 50% limit of data for the period T^i is set to guaranty that it has enough data to estimate the ratio between phases with less probability of error. This part of the algorithm will input all the missing samples where there is at least one adjacent sample and enough data on the three-phase time series to insert based on the ratio between phases. Equations (2) and (3) formulates the solution for a sample i of phase ϕ_a and, it can be similarly used for phases ϕ_b and ϕ_v.

For a given time stamp i where $X^i_{\phi_a} = null$ and $X^i_{\phi_b} \wedge X_{\phi^i_v} \neq null$.

$$X^i_{\phi_a} = \frac{1}{2}\left(\frac{\overline{X_{\phi_a}(T^i)}}{\overline{X_{\phi_b}(T^i)}} X^i_{\phi_b} + \frac{\overline{X_{\phi_a}(T^i)}}{\overline{X_{\phi_v}(T^i)}} X^i_{\phi_v}\right) \tag{2}$$

If only one adjacent sample exist $(X^i_{\phi_b} \vee X^i_{\phi_v}) \neq null$ then,

$$X^i_{\phi_a} = \frac{\overline{X_{\phi_a}(T^i)}}{\overline{X_{\phi_b}(T^i)}} X^i_{\phi_b} + \frac{\overline{X_{\phi_a}(T^i)}}{\overline{X_{\phi_v}(T^i)}} X^i_{\phi_v} \tag{3}$$

Finally, the last part will input data for more extended periods of consecutive three-phase missing values (periods of the day, days, weeks, and months), which is shown in Fig. 3 and described by the algorithm (1). Equation (4) is used to calculate the normalized standard day of the week curve $(S^{wd,\phi}_{std})$ where wd is the weekday, ϕ a specific phase and VD is the number of valid days (vd) for a specific weekday. It is essential to notice that X^{vd}_{ϕ} stands for all the samples of a valid day vd. A valid day is one with no missing values for any one of the three-phases. Furthermore, it is important to notice that if $VD < 3$ for any day of the week (wd), this means that there is not enough data to calculate the $S^{wd,\phi}_{std}$. Therefore, an alternative is to find another feeder time series with similar characteristics in the dataset.

$$S_{std}^{wd,\phi} = \frac{1}{VD}\sum_{d=1}^{VD}\frac{X_{\phi}^{vd} - min\left(X_{\phi}^{vd}\right)}{max\left(X_{\phi}^{vd}\right)} \quad (4)$$

As shown in Fig. 3, the second important information in order to use the $S_{std}^{wd,\phi}$ to fill parts of a day or whole days are the maximum and minimum values of each valid day (vd). The vectors are computed taking into account all the days in the time series, as described by algorithm 1. Additionally, in order to smooth any inconsistency, the moving average of two samples of the minimum and maximum vectors ζ and γ is used. Both pieces of information will be used to scale the NSSC curve to input on a specific day, as shown in (5) and (6).

Equations (4), (5) and (6) were used to fit the normalized standard weekday curve to a missing day on the dataset. Additionally, md stands for a missing day (more than 50% missing samples), γ is the vector of maximum values, and ζ is the vector of minimum values of each valid day of a specific phase of a given quantity of an MV feeder.

For a missing day (md) that is between valid days (vd) of same (wd),

$$X_{\phi}^{md} = \frac{1}{2}[(\gamma_{\phi}^{vd<md} + \gamma_{\phi}^{vd>md}) - (\zeta_{\phi}^{vd<md} + \zeta_{\phi}^{vd>md})]\cdot \\ \cdot S_{std}^{wd,\phi} + \frac{1}{2}(\zeta_{\phi}^{vd<md} + \zeta_{\phi}^{vd>md}) \quad (5)$$

if the missing day (md) is not between valid days (vd) of same (wd),

$$X_{\phi}^{md} = (\gamma_{\phi}^{vd_{closest}} - \zeta_{\phi}^{vd_{closest}})S_{std}^{wd,\phi} + \zeta_{\phi}^{vd_{closest}} \quad (6)$$

Where X_{ϕ}^{md} is a day with more than 50% of missing data, $\gamma_{\phi}^{vd<md}$ and $\zeta_{\phi}^{vd<md}$ are the maximum and minimum value, respectively, of a valid day of the same weekday before the missing day is filled. The $\gamma_{\phi}^{vd>md}$ and $\zeta_{\phi}^{vd>md}$ are the maximum and minimum value, respectively, of a valid day of the same weekday after the missing day is filled. If the missing day is not between two valid days, the closest one of same (wd) is used, as shown in (6).

Equations (4), (5) and (6) are also used to fill a period of the day (dawn, morning, afternoon, or night). The difference is that the result of the equations is sliced in a particular period of interest ($X_{\phi}^{md_{part}}$) before being inserted.

Figure 3 shows 21 days of one-phase current time series (light grey) with three days of missing data being filled (dashed dark black) and the values of the γ and ζ vectors. In this example, days 9 (Tuesday), 13 (Saturday), and 14 (Sunday) were missing from the middle week. Based on the maximum and minimum values of those same days from the week before and after, the missing days were filled using the normalized scaled standard day of the week curve (NSSC) obtained from the whole time series from that specific feeder.

One caveat of the imputation method proposed using the NSSC is that it must have at least three valid days for each weekday. It is possible that for a large amount of degradation, 60% or more, of the time series quantity (V, I, pf), there is not enough data to calculate the NSSC. The requirement of having at least three valid days was set empirically based on the analysis of the dataset used. Hence, an alternative is to use data from other feeders of the dataset to calculate the NSSC and apply it to the current feeder, as shown in the algorithm 1. The choice of the alternative feeder can be made considering the geographic region where each feeder is located or the characteristics of the majority of its consumers (households, industries, commercial buildings, etc.). In this work, the alternative feeder was chosen randomly on the database.

The final step is to apply another linear interpolation with $N_{samples} = \infty$ to take into account any missing sample that was not filled by the previous steps.

2.5 Probability density function of missing data

An essential step in studying the dataset and testing an imputation method is to know the characteristics of the missing data. The probability density function (PDF) describes the probability of a random variable to assume a given value and, in this case, would provide the likelihood of occurrence, duration, and the number of phases that were lost (Miller and Childers, 2004).

For the missing data, as stated previously, three PDFs must be obtained. The first one is the probability of a sample being missed. In this case, it was defined that it has a uniform probability. Hence, at any given time, the probability of a sample being lost is equal. Secondly is the probability of the type of a missing sample being of one, two, or three-phases. Finally, there is the PDF that describes the duration of the data lost, therefore, of losing one, two, fifty, or any given length. The last two probability density functions were determined empirically based on the histogram of occurrences of each type on the whole dataset. It is important to notice that the PDFs are different for each quantity (voltage, current, and power factor) (Murphy, 2012).

Knowing the PDFs, they can be used to tailor the imputation algorithm for optimal performance. Additionally, it can be used to degraded a valid time series at different levels, as described in section 2.6, and test the imputation method comparing with the original data.

2.6 Imputation method test methodology

The evaluation of the missing data imputation method was conducted in a sub dataset for each quantity (V, I, pf), where there was no outlier or missing data. The Fig. 4 shows the flowchart for testing the imputation method. As described, the first step is to find, for each quantity, the portion of a feeder in the dataset with no inconsistencies. Therefore, this subset is degraded in different levels by the probability density function (PDF) of missing data extracted from all the datasets, as discussed in 2.5. With the degraded time series, the imputation method proposed in 2.4 was applied and compared with the original data. This procedure was also done using the Naive approach in order to compare the two methods performance. The comparison of the two methods was conducted using three metrics: R-squared (R^2), mean absolute percentage error (MAPE), and root mean squared error (RMSE). Finally, it is important to mention that the time series

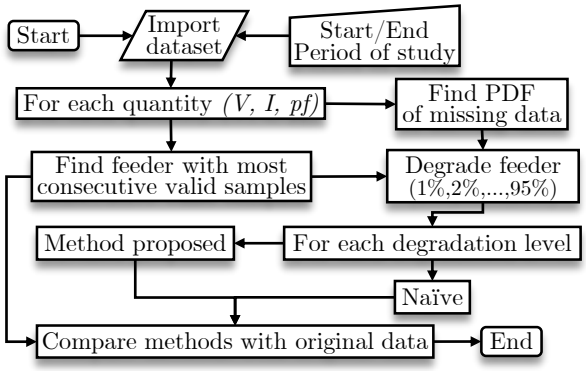

Figure 4. Flowchart of the imputation test methodology

were degraded from the following levels of data loss: $1\%, 2\%, 3\%, 4\%, 5\%, 10\%, 15\%, ..., 85\%, 90\%, 95\%$.

3. RESULTS AND DISCUSSION

3.1 Load transfer implication in bus voltage

A dependent sample t-test conducted the comparison of each one of the 115 MV buses. The comparison was between the average three-phase voltage during the load transfer of any of the related feeder and the average during regular operation (no load transfer). The results showed with 95% of confidence that there is no statistical difference for the three-phase voltage of the distribution substation during normal operation and load transfer of any of the related medium-voltage feeders.

3.2 Analysis of missing data

The analysis of missing data in the dataset of MV feeders can be done in two aspects. The first one regards the length or duration of consecutive missing samples, which indicates that a given attribute, for example, (V_ϕ), lost information for a sequence of timestamps. On the other hand, given that the quantities studied are the combination of three time series, an important aspect is the number of phases that were lost in a specific timestamp. Figure 5 shows the percentage of occurrences in the dataset of each type of data lost, whereas Fig. 6 shows the percentage of occurrences of each length of consecutive data samples lost for the period between January the 1st and December the 31st of 2019.

Most missing values, 73.99% for voltage, 90.53% for current, and 82.75% for power factor, comprehend the loss of all three-phases. However, the dataset still has missing values of only one and two-phases: 26.01% for voltage, 9.47% for current, and 17.25%, as shown in Fig. 5.

Regarding the length of consecutive missing samples, the majority are of one sample. In the dataset of MV feeders, 95% of the occurrences were up to a duration of four samples for voltage, up to nine samples for current and up to thirty-five for power factor, hence, less than two hours and 55 min. Although most of the occurrences are far from days of duration, it is essential to notice that for a given feeder, one occurrence of consecutive three-phase loss of $3x10^4$ samples is sufficient to compromise the analysis of the feeder with months of missing values.

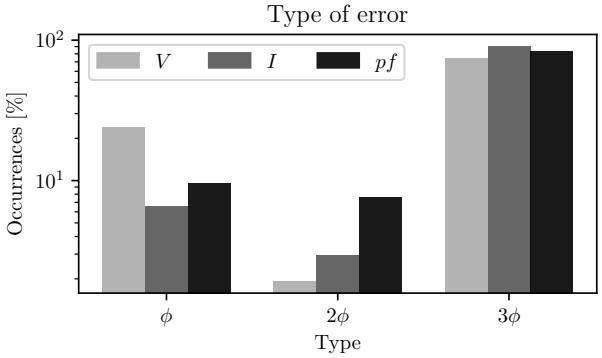

Figure 5. Percentage of occurrences of one, two and three-phase data loss.

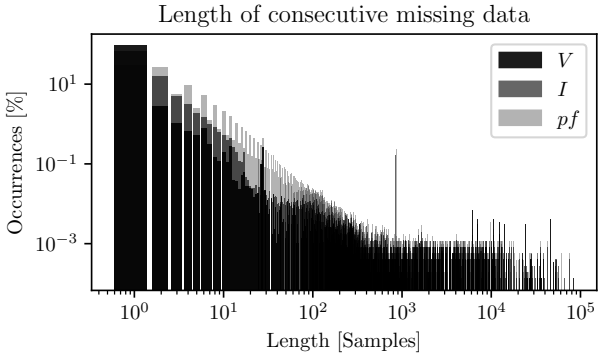

Figure 6. Percentage of occurrences of each consecutive missing samples length.

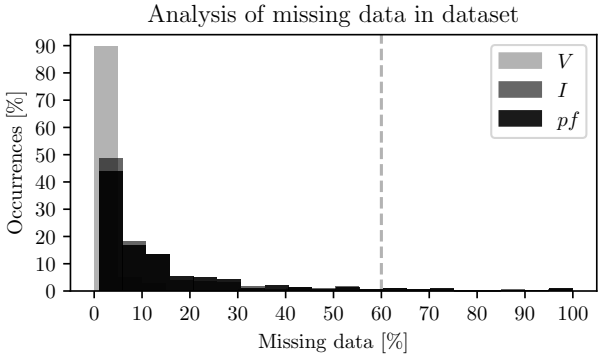

Figure 7. Percentage of missing values per feeder in the dataset.

Figure 7 shows the percentage of loss for each feeder in the dataset. For the period of study, 98.91% of feeders lost less than 60% of the voltage information, 96.95% of feeders lost less than 60% of the current information, and 95.39% of feeders lost less than 60% of the power factor information.

3.3 Imputation method

The results shown in this section were obtained after 20 executions of the method on each degradation level, as shown in Fig. 4. Figures 8, 9 and 10 show the result for the method proposed in section 2.4 using real feeder data

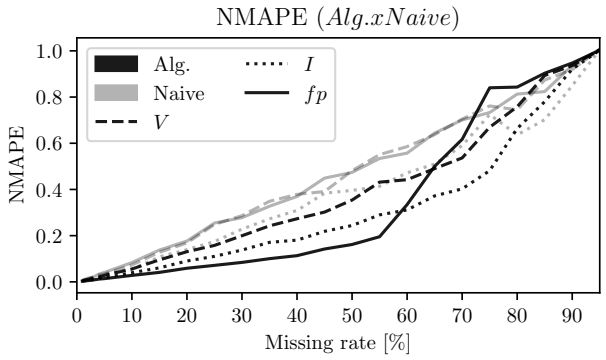

Figure 8. Evaluation of the method proposed for different degradation levels using NMAPE.

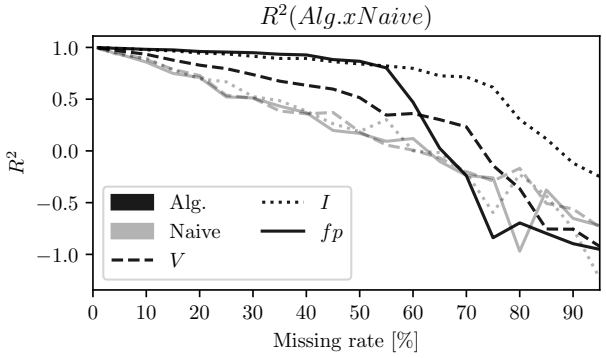

Figure 9. Evaluation of the method proposed for different degradation levels using R^2.

and tested as described in section 2.6. Figure 11 shows the statistical differences between the method and the Naive approach using a dependent sample t-test with 5% α. The procedure discussed previously of missing data imputation has better performance than the Naive approach for most of the degradation levels tested. However, for more than 60% of missing data, the performance starts to degrade rapidly. For the power factor, the point of no statistical difference between the methods stars at 65% whereas for voltage, it is 75%. For current, the method is statistically better than the Naive for any of the values tested. The normalized version of the metrics discussed in section 2.6 was an alternative to accommodate all the quantities results on the same graph for analysis.

Figure 12 shows six months of a three-phase current time series. The original data has no outliers or missing values. The 25% degraded version has long periods of three-phase and one-phase data loss. This curve was obtained by the method discussed in 2.6. The last two graphs in Fig. 12 shows the results of the proposed method in 2.4 and the Naive approach of missing data imputation.

4. CONCLUSIONS

In this study, a method of preprocessing and missing sample imputation for medium voltage feeders is proposed based on the analysis of the missing values after the outlier and load transfer removal. It was verified that the information of the medium voltage feeders sampled at the

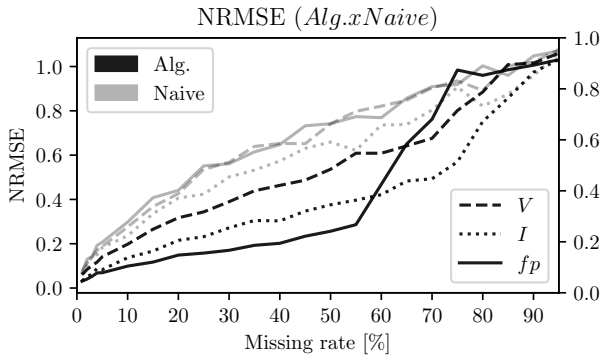

Figure 10. Evaluation of the method proposed for different degradation levels using NRMSE.

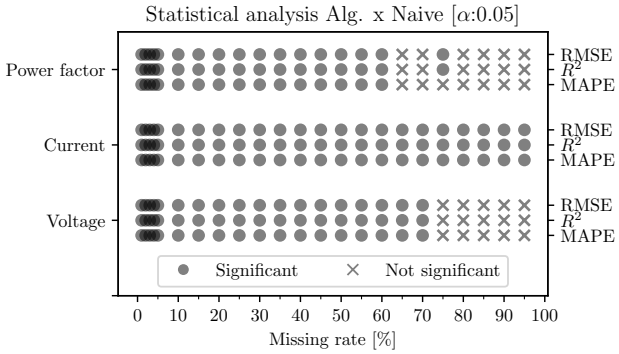

Figure 11. Statistical analysis: performance of the algorithm proposed and the Naive approach.

substations relays must be treated before being used on the distribution network planning as it may still be impossible to avoid data incompleteness or with the absence of outliers. In this context, the process of synchronizing the samples is of extreme importance in order to perform operations with the voltage, current, and power factor of each feeder. Furthermore, it provides the capability to analyze load transfers and correlations between feeders. A three-part process is proposed to remove the outliers. In the first part, the maneuvers are removed based on the load maneuvers dataset. Secondly, the its the removal of other samples that do not respect the physical and or theoretical constraints of the system. Finally, its applied a statistical method based on the median absolute deviation around the moving median to contemplate the seasonality of the feeder.

The missing values analysis showed that most of the missing samples were of three-phase nature. However, it still exists a significant percentage of one or two-phase voltages that were addressed by the ratio between phases. Regarding the length of consecutive missing values, the majority is of less than two hours and 55 min. For these samples that could not be filled by the ration between phases, the linear interpolation was used. For more extensive periods of three-phase missing values were imputed by the normalized scaled standard weekday curve (NSSC). This method utilizes the correlation between the quantities and the periods of the days and weekdays.

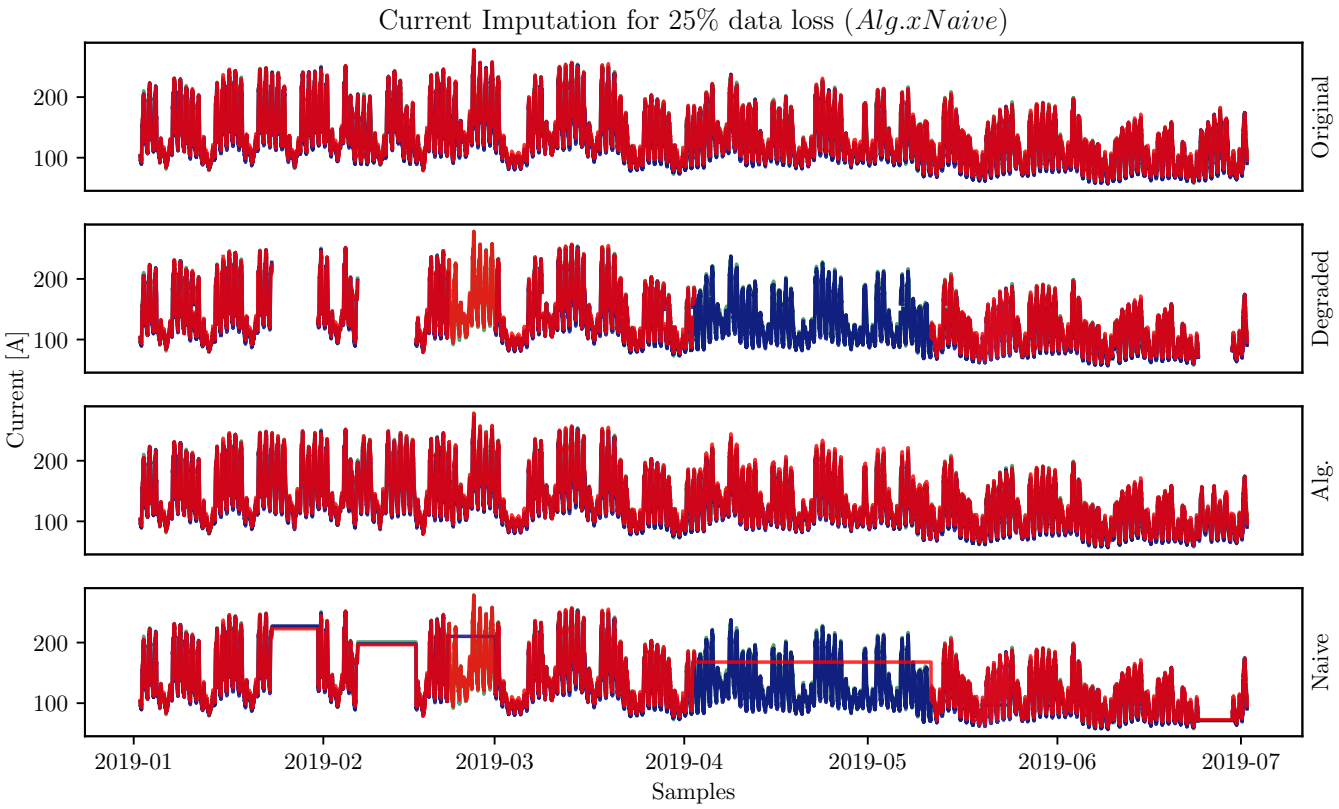

Figure 12. Example of an current time series with all the samples and its version degraded by 25%. Its all so shown the imputation method proposed and the Naive results. The curves in red, green and blue, shows the phases ϕ_a, ϕ_b, and ϕ_v respectively.

The method proposed was compared with the Naive approach and shows promising results that were statistically significant for up to 60% of feeder degradation.

REFERENCES

Gonen, T. (2007). *Electric Power Distribution System Engineering*. CRC Press.

Han, J., Kamber, M., and Pei, J. (2012). *Data mining: concepts and techniques*. Morgan Kaufmann.

Jadhav, A., Pramod, D., and Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33(10), 913–933. doi:10.1080/08839514.2019.1637138. URL `https://doi.org/10.1080/08839514.2019.1637138`.

LEYS, C., KLEIN, O., BERNARD, P., and LICATA, L. (2019). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation a round the median. *Journal of Experimental Social Psychology*, 764–766.

Miller, S. and Childers, D. (2004). *Probability and random processes: with applications to signal processing and communications*. Elsevier Academic Press.

Muñoz-Delgado, G., Contreras, J., and Arroyo, J. (2018). *Distribution System Expansion Planning*, 1–39. doi:10.1007/978-981-10-7056-3_1.

Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. The MIT Press, 1 edition.

Peppanen, J., Zhang, X., Grijalva, S., and Reno, M.J. (2016). Handling bad or missing smart meter data through advanced data imputation. *2016 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2016*, 0–4. doi:10.1109/ISGT.2016.7781213.

Saunders, J.A., Morrow-Howell, N., Spitznagel, E., Doré, P., Proctor, E.K., and Pescarino, R. (2006). Imputing missing data: A comparison of methods for social work researchers. *Social Work Research*, 30(1), 19–31. doi:10.1093/swr/30.1.19.

Shier, R. (2004). Statistics: 1.1 paired t-tests. *Mathematics Learning Support Centre*.

Vargas, E.L. (2015). *Planejamento da expansão do sistema de. Distribuição através da simulação de alternativas e análise mul ticritério*. Master's thesis, Universidade federal de Santa Maria, Santa Maria.

Wen-Chih Yang, W.T.H. (2011). An enhanced load transfer scheme for power distribution systems connected with distributed generation sources. *WSEAS TRANSACTIONS on CIRCUITS and SYSTEMS*.