UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
DEPARTAMENTO DE PÓS-GRADUAÇÃO
EM ENGENHARIA ELÉTRICA

Missing Data Analysis and Imputation Method for Medium Voltage Distribution Network Feeders

João Marcus Ramos Bacalhau

Vitória

2020

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
DEPARTAMENTO DE PÓS-GRADUAÇÃO
EM ENGENHARIA ELÉTRICA

João Marcus Ramos Bacalhau

Missing Data Analysis and Imputation Method for Medium Voltage Distribution Network Feeders

A dissertation submitted as partial fulfillment of the requirements for the degree of Master's in Electrical Engineering at the Universidade Federal do Espírito Santo.

Supervisor: PhD. Jussara Farias Fardin

Vitória

2020

Ficha catalográfica disponibilizada pelo Sistema Integrado de
Bibliotecas - SIBI/UFES e elaborada pelo autor

R175
m

Ramos Bacalhau, João Marcus, 1992-
Missing data analysis and imputation method for medium voltage distribution network feeders / João Marcus Ramos Bacalhau. - 2020.
92 f. : il.

Orientadora: Jussara Farias Fardin.
Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Network expansion. 2. Data imputation. 3. Time series data. 4. Distribution system planning. 5. Data analysis. 6. Incomplete data. I. Farias Fardin, Jussara. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 621.3

João Marcus Ramos Bacalhau

Missing Data Analysis and Imputation Method for Medium Voltage Distribution Network Feeders

Dissertation submitted as partial fulfillment of the requirements for obtaining a Master's Degree in Electrical Engineering at the Universidade Federal do Espírito Santo.

Approved on 05/11/2020, Vitória-ES.

PhD. Jussara Farias Fardin
Universidade Federal do Espírito Santo.
Supervisor

PhD. Lucas Frizera Encarnação
Universidade Federal do Espírito Santo.
Examiner

PhD. Daniel Cruz Cavalieri
Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo
Examiner

Vitória, ES, Brazil

November 2020

Contents

List of Figures

Abbreviations

R^2	R-squared
AR	Autoregressive
CSV	Comma Separated Values
DSO	Distribution System Operator
HV	High Voltage
I	Current
LOCF	Last Observation Carried Forward
MAPE	Mean Absolute Percentage Error
MV	Medium Voltage
NSSC	Normalized Scaled Standard day of the week Curve
PDF	Probability Density Function
PDFs	Probability Density Functions
pf	Power Factor
RMSE	Root Mean Squared Error
V	Voltage

Nomenclature

$E_{active/reactive}$	Active and reactive energy
M_j	Moving mean with windows of j samples
$Max_{threshold}$	Superior limit for the sample to be considered an outlier
$Min_{threshold}$	Inferior limit for the sample to be considered an outlier
N	Number of standard deviations
$N_{samples}$	Limit to the number of samples that can be interpolated
$P_{active/reactive}$	Active and reactive power
$S_{std}^{wd,\phi}$	Normalized standard weekday curve of weekday wd and a given phase ϕ
T_i	Period of analysis where the sample i is located
VD	Number of valid days for a given weekday
V_ϕ	A given phase of the voltage time series
$X_{\phi_a}^i$	Sample of phase ϕ_a at timestamp i
$X_{\phi_b}^i$	Sample of phase ϕ_b at timestamp i
$X_{\phi_v}^i$	Sample of phase ϕ_v at timestamp i
$X_\phi^{md_{part}}$	Period of the day of a missing day md of a given phase ϕ
X_ϕ^{md}	All samples of a missing day md of a given phase ϕ
X_ϕ^{vd}	All samples of a valid day vd of a given phase ϕ
γ	Minimum day values vector
$\gamma_\phi^{vd<md}$	Maximum value of a valid day of same weekday before the missing day
$\gamma_\phi^{vd>md}$	Maximum value of a valid day of same weekday after the missing day
μ	Mean
ϕ_a	Phase A of a given quantity
ϕ_b	Phase B of a given quantity
ϕ_v	Phase V of a given quantity
ζ	Maximum day values vector
$\zeta_\phi^{vd<md}$	Minimum value of a valid day of same weekday before the missing day
$\zeta_\phi^{vd>md}$	Minimum value of a valid day of same weekday after the missing day
i	Timestamp of interest

$i+1$	Subsequent timestamp
$i-1$	Previous timestamp
j	Number of samples of the moving window
md	Day with more than 50% of missing samples
vd	Valid day. A day with no missing samples
wd	Weekday

Acknowledgement

First, I would like to recognize the importance of two strong and beautiful women, my sister Anna and my girlfriend Isabella. My dear sister, I honestly can't stop finding in you, such an accomplished scientist, the inspiration to reach further steps into my career. Thank you for the best-unbiased pieces of advice that anyone could have wished for. Isabella, honey, thank you so much for all your support, without which I would have stopped these studies a long time ago. You were the one who has supported me the most and had to put up with my stresses and complaints for the past three years of study. Your dedication and commitment to your work amaze me every day; you are definitely a role model. To my parents, thank you for putting up with me being sat in the computer for hours on end and for providing guidance through this research.

To my colleagues at University, thank you for sharing so many extraordinary moments and for helping me make the most of the courses that we took. To my work colleagues, thank you for the conversations that were vital in inspiring me. To my closest friends, thank you for all the support and for all the funny moments that made life much more enjoyable.

Finally, and just as importantly, I would like to thanks my supervisor, Ph.D. Jussara Fardin, for providing guidance and feedback without which this work would not be complete. Thank you, professor, for giving me the freedom to choose my path and independence to fulfill the credits as I thought it would fit my dream.

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

Abstract

The energy sector's investment aims to ensure a continuous, reliable, and quality supply of electrical energy imposed by the electricity regulatory agency with maximum economic-financial balance. This dissertation 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, the use of the standard curve for each day of the week is proposed. 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.

Resumo

O investimento no setor de energia visa garantir um fornecimento contínuo, confiável e de qualidade de energia elétrica impostos pela agência reguladora de eletricidade com o máximo equilíbrio econômico-financeiro. Esta dissertação discute os desafios do processamento de dados de alimentadores de distribuição de média tensão para uso no planejamento de redes de distribuição. A análise dos dados perdidos e dos *outliers* é feita nas séries temporais trifásicas das medidas de tensão, corrente e fator de potência de 459 alimentadores reais. Além disso, é proposto um método de pré-processamento e imputação de dados ausentes, usando a característica desequilibrada das fases, interpolação e a curva padrão normalizada do dia da semana ajustada. Os resultados obtidos mostram que a maior parte dos dados perdidos são trifásicos, contudo, com uma quantidade significativa de amostras monofásicas e bifásicas perdidas que podem ser preenchidas utilizando a proporção entre fases. Portanto o desafio se torna preencher múltiplas semanas de dados trifásicos perdidos e para isso um método que utiliza a curva padrão do dia da semana é proposto. O método proposto se mostra como alternativa promissora para a imputação de dados em alimentadores de média tensão. A técnica é testada utilizando dados reais de alimentadores degradados pela própria função densidade probabilidade de dados perdidos e comparada com um método ingênuo.

Palavras-chave: Expansão de rede; Planejamento do sistema de distribuição; Imputação de dados; Alimentadores; Análise de dados; Valores ausentes; Dados incompletos; Imputação; Dados de séries temporais.

1 Introduction

1.1 Power distribution network planning

Distribution networks are the last mile on the energy delivery from the generators to the end-users. Typically, following a radial topology, 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, load characteristics variability, and 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 system's capability, maintaining a continuous, reliable, and quality supply of electrical energy (Vargas, 2015). Specifically, in Brazil, the electrical energy distribution is regulated by the government through, but not limited to, the Distribution Code (PRODIST). The PRODIST states the conditions and responsibilities for the expansion planning, customer access conditions, operation, measurement, indicators, technical losses, and power quality. Furthermore, it also defines procedures and establishes criteria for the information exchange between the parties (Aneel, 2017).

The distribution planning relies on the quality of the information and availability to make decisions on the sector. The electrical quantities historical data combined with market projections are used to forecast years ahead of the energy demand of regions with its substations and feeders. These electrical quantities projections are the input data for a power distribution system simulation that will indicate weak points of the power network. Hence, the utility company reinforces its assets in preparation for the demand increase. Therefore, the lack of reliable data directly impacts the strategic objectives of electricity companies and the efficiency of medium and long-term investments.

1.2 Data analysis and missing data imputation

The historical information analysis is a very powerful tool to discover trends and patterns in businesses. Since the invention of the firsts calculators in the 17th century,

our capability to process data has been improved considerably. The firsts data storages, as we know today, were created in the 1960s with the firsts modern advanced database management systems been available in the 1980s. The processing of information, as we understand and use today, was also implemented in the late 80s. However, only in the mid-2000s with the reduction in costs of data acquisition and with machine learning algorithms moving from laboratories to industries, we started to contemplate what this technology is capable of providing. Regarding the theory behind these techniques that are currently being used, most were developed decades ago. However, only with the progress in computation power in the recent 20 years, we were able to apply it to large datasets across many different areas. It is important to notice that the computation leap also improved our capability to collect data more precisely, with minimum loss of information and in higher sample frequencies (Han et al., 2012; Google, 2020).

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 contributors to the process are the equipment failure that alters measurements and weather conditions that prevents the information transfer to the DSO. In conjunction with the stop of equipment for preventive maintenance and load transfers (specifically in the context of system planning), these facts cause outliers and vacancies to appear on the dataset.

Since inappropriate treatment of missing values may cause incorrect results in data mining, the missing value imputation problem has become the focus of the incomplete data analysis in opposition to sample deletion. Several imputation methods have been proposed for time series data, 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 neighbour, 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., 2006b).

In this work, an imputation method is proposed based on the correlation of the time series studied and the weekdays. Furthermore, the correlation between time series (three-

phase systems) and interpolation is also used to fill missing values.

The method proposed in this work will be compared with a Naive approach for filling missing samples. Naive approaches for imputing missing values normally tend to have a poor performance for situations of high complexity. On the other hand, these techniques are easy to implement, do not demand a large amount of previous data to be trained, and can be executed quickly. One Naive imputation method is the Hot Deck by Last Observation Carried Forward (LOCF). In this method, the last value observed in the time series is carried up to a point where another sample is found. The LOCF is the chosen technique to compare the proposed method as commonly used in the literature and has a suitable performance for most of the missing values encountered in the studied dataset, which is up to three hours. The result of the analysis of length and type of missing data values is presented in section 3.2.

1.3 Literature review

There are several studies in the literature and many different approaches regarding data imputation. The main keywords used in this literature review were missing value, multiple imputations, missing data, missing data imputation, data analysis, data imputation methods, missing value imputation, time series imputation, incomplete data, imputation, and time series data. The research for related work was focused on the following search engines: IEEE Xplore, Web of Science, Scopus, ScienceDirect, and Google Scholar. Although any work related to the general approach to missing value imputation was considered during the literature review (Bashir and Wei, 2016; Saunders et al., 2006a), the focus was on finding related work of imputation methods using real-world data, and more specifically, the imputation of missing values on MV distribution power network feeders.

In Zhang et al., the authors propose a combined model and compares it with line averaging, linear regression, correlation analysis, and clustering. Furthermore, the authors present a power flow analysis solution for the grid buses where there is no measurement device, hence, no historical data is available. The one month dataset used was provided by a power supply company with a sample period of one hour, and a test was made

considering the loss of one day. Additionally, for the power flow simulation, the authors used the IEEE 16 bus distribution system. The authors concluded that the combined method based on least square theory is efficient for missing load data where measuring equipment is placed. Finally, It is stated that the use of power flow analysis can be an alternate for the points with no measurement.

The paper presented by Gheorghe et al. (2009) two imputation methods, the k-Nearest Neighbors (kNNs) and Clustering, are considered in the treatment of missing data on electric distribution networks. The analysis is conducted with three distinct lengths of consecutive missing values: three, five, and nine samples. For the application of the kNNs method, the authors separated the dataset into a training set to determine the number of neighbours and a test set for the proposed method validation. From the obtained result analysis, it can be observed that the largest errors correspond to the period with the nine consecutive missing values in both methods. Using MAPE it was concluded that the method of clustering the typical load profiles is more robust than KNNs method and requires less work.

In Chang et al. (2015) a new imputation method to fill up the missing slots in the time series data is proposed taking advantage of the temporal information. By presenting explicitly local time indices and non-missing values to the least squares support vector machine (LSSVM), the authors affirm that missing values can be computed efficiently and naturally. The method is tested by experiments done on three real-world datasets: the Poland dataset with 1500 records of electricity load, the Laser dataset (Pulse intensity) with 5700 records, and the Sunspot dataset with a year measurements in a total of 288 records. The RMSE was computed, and the method was compared with the mean, Hot Deck, and Autoregressive (AR). The degraded dataset was generated by randomly deleting samples in the following missing rates: 1%, 5%, 10%, 15%, 20%, and 25%. Although there were some missing rates for the sunspot dataset were the proposed method had superior performance, the AR had an overall better performance. In the Laser dataset, the AR model was the best alternative and, finally, the proposed method performed better in the Poland dataset.

The use a bottom-up load curve simulation with multiple layers called MOSAIC is

proposed by Kong et al. (2017). The algorithm evaluates the load curve for each type of customer (residential, solar photovoltaic, wind power producer, etc.) and aggregates them up to the desired scale, for example the feeder. As stated by the authors this requires having a significant amount of information on the study area. When information is missing, the simulator uses instead default values, first at block level, second at project level if needed. This multi-layer organization makes it possible to simulate a power distribution network area where some data are missing. Nevertheless, the quality of the result will depend on how precise the area is described. The work is more related with the forecast of the curve itself rather than imputing missing data, nevertheless, It is an approach that can surely be used to substitute the feeder measurement in the substation in case of long periods of data loss.

The study conducted by Mccoy et al. (2018) a Variational Autoencoder, a recent Deep-learning technique used for imputing data on images, is proposed to fill missing values on a simulated milling circuit dataset. The authors also compares the technique with mean replacement and by principal component analysis (PCA) imputation. Furthermore, the performance was verified using RMSE in two levels of degradation 20% (Light) and 90% (Heavy).

In Razavi-Far et al. (2020) the authors propose a missing data imputation technique that can handle both numerical and categorical features. The two novel missing data imputation techniques proposed are based on the k-Nearest Neighbour and Expectation–Maximization algorithms. The aim is to learn both similarities between the records (local similarity) and correlation among the features (global similarity) within the dataset. The work compares eight imputations techniques, including the two proposed, in twenty-one different datasets. The results obtained shows that the proposed combined algorithm is superior than its competitors.

In Huyghues-Beaufond et al. (2020) the authors focuses on short-term load forecasting of distribution feeders. Nevertheless, it is proposed a hybrid automatic outliers cleaning procedure which is suitable for large datasets. Additionally, the authors considered three primary imputation techniques for the estimation of missing observations: Unconditional Mean, Hot Deck based on K Nearest Neighbor, and Kalman smoothing, recommend-

ing the Hot Deck (k-NN) because of its better performance. Having the 342 MV feeders preprocessed by the outlier detection and missing values imputation techniques they modelled the MV feeders data with feed-forward deep neural networks to forecast up to 24 steps-ahead.

Additionally, in Jadhav et al. (2019) and Amiri and Jensen (2016) the authors makes a thorough review of different methods for imputing missing data. It is important to notice that the last one is focused on general algorithms, hence, not intended to be used in a specific field of study.

The related work demonstrates that extensive literature regarding different methods of dealing with missing data exists. The idea of testing multiple degradation levels in the present work was similarly used by Chang et al. (2015), Mccoy et al. (2018), and Gheorghe et al. (2009). The combined model presented by Zhang et al. was a reassurance that multiple methods used to specific parts of the problem would be an alternative to a single method to solve all inconsistencies. The performance metrics similarly used by Gheorghe et al. (2009), Mccoy et al. (2018), Razavi-Far et al. (2020), and Huyghues-Beaufond et al. (2020) were an inspiration for the normalized metrics used to compare the proposed imputation method and the Naive approach. In Huyghues-Beaufond et al. (2020) the importance of imputation methods is shown as the first step in a forecast algorithm using neural networks. Finally, Kong et al. (2017) shows that the data loss can be prevented using multiple sources of data on the distribution network and the correlation between them.

This work presents another approach to the missing data imputation problem for MV feeders using a combined model with the Normalized Scaled Standard day of the week Curve (NSSC), phase correlation, and linear interpolation. Furthermore, a different test method is proposed where slices of data with no missing values found in the dataset are degraded by 23 different levels (1% to 95%) based on the Probability Density Function (PDF) of missing values of real feeders. The degraded versions are reconstructed by the method proposed and compared with the reconstruction made by the Naive approach. It is important to mention that the most common metrics in the literature are used to assess the method's performance. Additionally the method is simple to implement, does

not require the operator's interference during its execution, and has high interpretability. The techniques used are the summary and extension of the experiences acquired by investigating the system measurement's characteristics and its correlation with time and its electrical quantities.

1.4 Dataset

This research utilizes a dataset that consists of nine time series for each of the 459 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. Additionally, the dataset is organized in Comma Separated Values (CSV) files with 12,1 Gigabytes in total size. Figure 1 shows the nine time series of a MV feeder after the preprocessing and imputation method proposed being applied.

Timestamp	$I\Phi_a$	$I\Phi_b$	$I\Phi_v$	$V\Phi_a$	$V\Phi_b$	$V\Phi_v$	$fp\Phi_a$	$fp\Phi_b$	$fp\Phi_v$
06/01/2019 15:25:00	71,90	78,93	74,56	11,69	11,65	11,63	0,99	0,98	0,98
06/01/2019 15:30:00	73,15	80,34	78,15	11,61	11,55	11,52	0,99	0,98	0,98
06/01/2019 15:35:00	70,18	77,06	75,34	11,59	11,55	11,51	0,99	0,98	0,98
06/01/2019 15:40:00	73,30	79,71	77,52	11,66	11,58	11,58	0,99	0,98	0,98
06/01/2019 15:45:00	73,30	79,71	77,52	11,56	11,52	11,49	0,99	0,98	0,98
06/01/2019 15:50:00	73,30	77,68	75,65	11,58	11,53	11,49	0,99	0,99	0,98
06/01/2019 15:55:00	73,30	80,18	78,31	11,69	11,65	11,63	0,99	0,99	0,98
06/01/2019 16:00:00	70,33	78,15	75,96	11,69	11,65	11,61	0,99	0,98	0,97

Figure 1: Sample data of one MV feeder showing the nine time series.

Similarly to other utilities, the relay's portfolio 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 where 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 involved.

In this work, 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.

1.5 Context and contribution

In this dissertation, 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 time series of real feeders. Furthermore, it is proposed a method of preprocessing with outlier removal, missing data imputation using the unbalanced characteristic between phases, time series interpolation, and the NSSC.

The proposed method is tested using real data degraded by 23 different levels (1% to 95%) based on the PDF of missing values of real feeders. The preprocessing and imputation method proposed are 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.

1.6 Document structure

This work is divided into three chapters. The first one is an introduction to the main topics regarding this dissertation, including the power system planning, the dataset used, and other related research. In the second chapter, the methodology is discussed. The importance of synchronization and outlier removal is clarified, and the imputation method algorithm with the related equations is presented. This chapter is also where the method to test the algorithm is described alongside the procedure to obtain the missing data's probability density function. Finally, on the third one, the outcomes obtained are shown and discussed. Furthermore, a summary of the results is presented alongside a proposal for future work. At the end of this dissertation, the appendices show, as an

example, the results obtained from a single execution of the test procedures described.

1.7 General Objective

The main objective of the research is to analyze missing data in the time series of voltage, current, and power factor of medium voltage feeders and to propose an imputation method to fill all vacancies.

1.8 Specific Objectives

The following topics are the specific steps intended to be followed in order to reach the main objective proposed:

- Analyse the characteristics of the missing data in the context of this research;
- Propose a method to synchronize all time series and reuse data to pre-fill missing values;
- Propose a method of testing an imputation method based on real time series and the PDF of missing data;
- Compare the result with a well-known method of imputation using standard metrics.

2 Methods

In this chapter, all the main steps of the proposed imputation algorithm are described. The synchronization and its importance, the outlier removal and its impact on the time series, and the imputation method with related equations are detailed and explained. Furthermore, the load transfer between feeders effect is addressed, and the process of testing the algorithm and obtaining the vacancies PDF is explained. Figure 2 shows the sequence of steps from the raw dataset up to the performance comparison between the method proposed and the Naive approach. It is important to notice that every step shown in Figure 2 is conducted by a series of python scripts, however, it does not require the operator intervention besides the start of each major component (synchronization, outlier removal, and imputation).

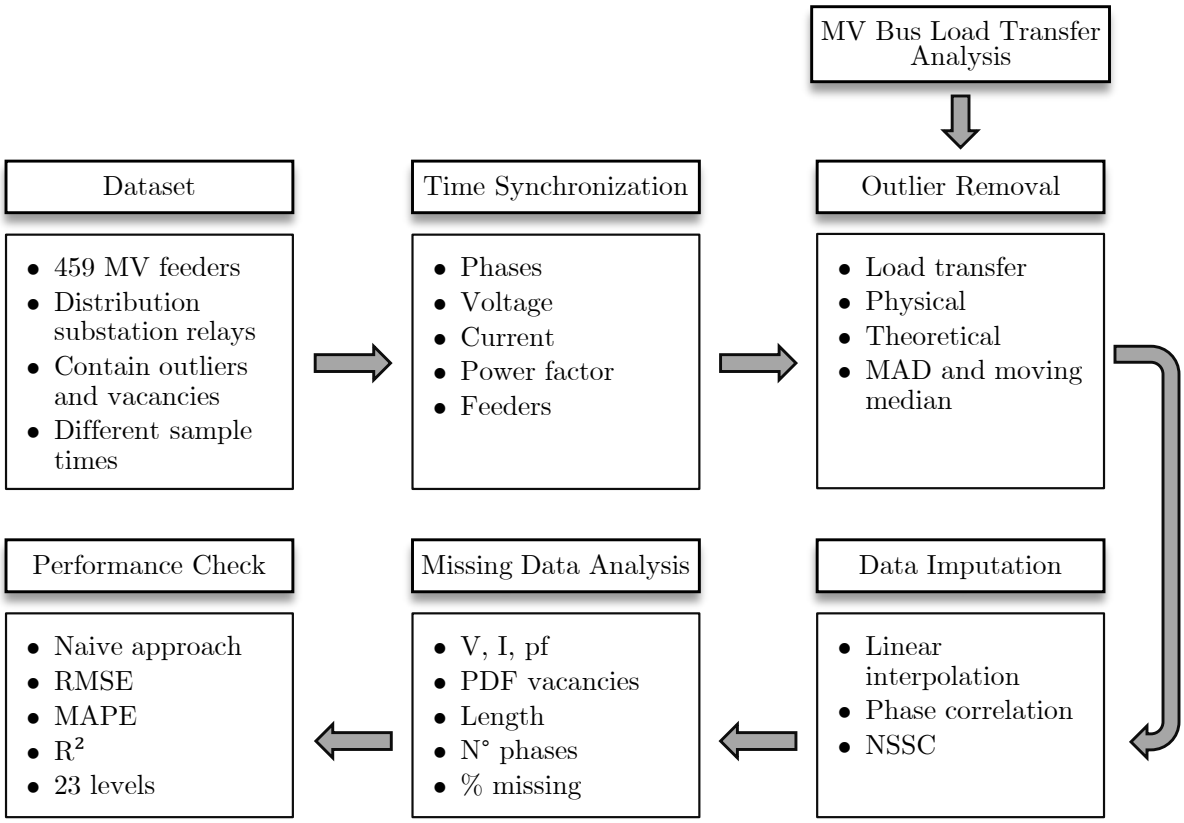

Figure 2: A flowchart of the proposed method with additional procedures. The auxiliary procedures are the performance check, the discussion regarding the load transfer impact on the bus voltage, and the missing data analysis. Auxiliary procedures described are not intended to be run in a production environment, however, they are vital to assess the performance of the 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 (ϕ_a, ϕ_b, ϕ_v) of the same quantity, between quantities (V, I, pf) of the same feeder, and between feeders, provides many advantages. The first one being the ability to combine all nine time series, the three-phase voltage, current, and power factor of each feeder to calculate the secondary quantities ($P_{active/reactive}$, $E_{active/reactive}$). Figure 3 shows an example of a distribution substation with a High Voltage (HV) bus, two MV buses (MV 1 and MV 2), two transformers (T1 and T2), and its feeders FD01 to FD04. Each feeder is represented by a color with its customers (coloured squares), and the normally open reclosers used for load transfer.

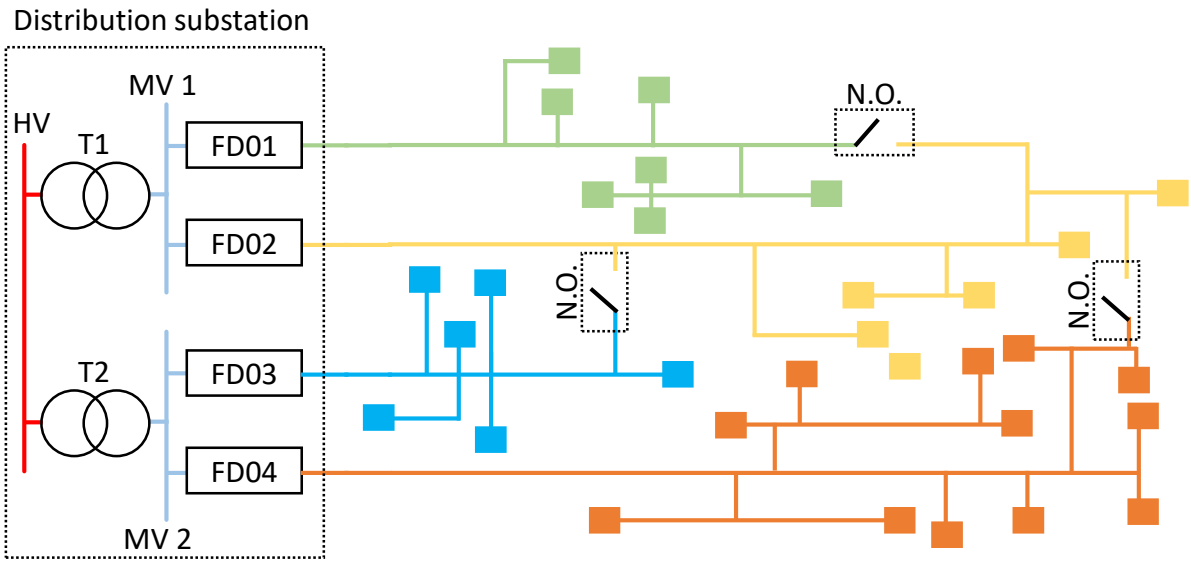

Figure 3: Example of a power distribution network.

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 substation's transformers quantities by the sum of all feeders. For example, the electrical quantities estimation of the substation transformer T1, shown in Figure 3, can be obtained by the time series sum of FD01 and FD02.

Finally, in the dataset, there are two different time sampling periods. The first one, which is predominant, is 5 min, and the other one is 1 min. To make the sampling period uniform between feeders those which were sampled at 1 min were downsampled to 5 min.

Hence, this process also reduces the size of the dataset. It is essential to notice that the sample period of a feeder can change. For example, a relay that is sampling at 5 min periods can be replaced for a new model that samples at 1 min. Therefore, this process is also used to standardize the sampling period of a given feeder during the period of study.

In order to synchronize the samples, 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. One caveat is that samples $i - 1$ and $i + 1$ should not be more than 2 minutes and 30 seconds distant of sample i. This assumptions will provide the ability to retain a large number of samples that would be otherwise discarded as they were not sampled exactly 5 minutes apart. For example, the sample 01/01/2019 00:15:00 would be valid if recorded in the database between 00:12:30 and 00:17:30.

Another important information is the period of study also called the period of interest. The dataset used has samples from 01/01/2019 00:00 to 31/12/2019 23:55, therefore, it is expected for every feeder to have samples from 01/01/2019 00:00 to the end of the period of study (31/12/2019 23:55) in intervals of 5 min (00:00, 00:05, 00:10, 00:15, etc.). The fact that all feeders starts at 01/01/2019 00:00 will provide the ability, as explained, to perform computations between them.

Figure 4 shows a slice of the three-phase voltage time series and the process of synchronizing and downsampling. In the example, phases ϕ_a and ϕ_v of sample one were shifted by one second to fit the expected standard start for the feeder at 00:00:00. Additionally, phases ϕ_a and ϕ_v from sample 5 were used to fill the missing data from sample 6 which is less than 2:30min distant from each other, and phase ϕ_b from sample two was used to fill sample one using the sample principle. It is important to notice that samples three and four are too distant to be used to fill the timestamp 00:05:00. Regarding the sample 6, sample 5 is a better candidate (closer), and the sample 2 is a better candidate (closer) to fill the missing phase of sample one. Therefore, samples three and four were discarded.

On the bottom of Figure 4, 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 (section 2.6), the imputation

method proposed will insert the missing samples. Nevertheless, the three main steps in the synchronization process are: synchronizing the start of the time series, utilize samples $i+1$ and $i-1$ to fill missing data in sample i, and downsampling. As stated before, 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.

Sample	Timestamp	V_{ϕ_a}	V_{ϕ_b}	V_{ϕ_v}
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	V_{ϕ_a}	V_{ϕ_b}	V_{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 4: Raw data synchronization and downsampling for three-phase voltage.

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 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 Figure 6, had the beginning of its operation (commissioning) in sample 1250, therefore, all the samples before that timestamp were 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 Figure 6 by the example of a one-phase current time series and in Figure 5 by the flowchart. 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.

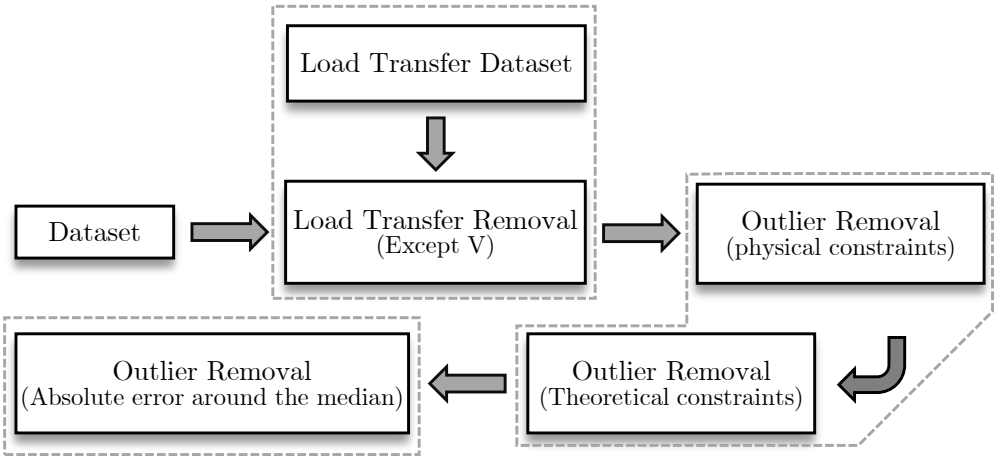

Figure 5: Flowchart of the steps in the outlier removal procedure.

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. 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. Another important topic is about recurrent load transfers made to improve the system operation. In some situations, it is possible that the DSO repeatedly maneuvers a feeder in a certain period of the day or season to reduce technical losses or improve the system stability. In this case, this type of maneuvers should not be recorded in the load transfer dataset as they are not considered outliers. In Figure 6 around sample 1500, the feeder receives load, and therefore this period is marked as not valid and removed.

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, hence, should be removed as outliers. For the case shown in Figure 6, the physical capacity of the cable is 220A (shown by the horizontal dashed line), therefore, samples 2000 and 2500 where marked and removed as outliers.

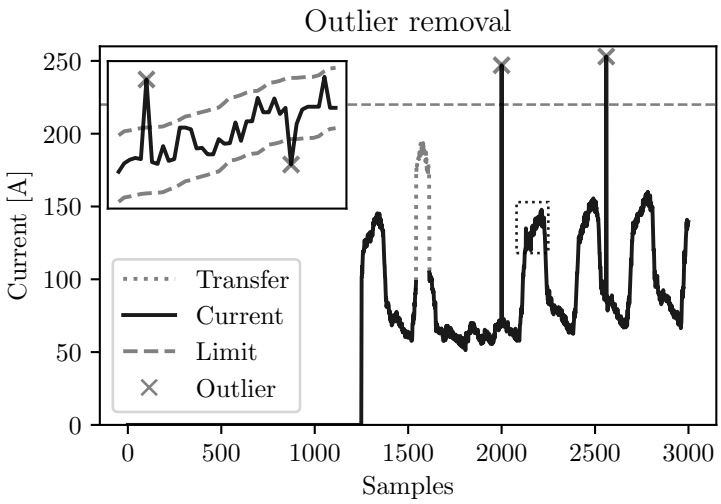

Figure 6: Example of outlier removal for current.

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. Furthermore, the authors propose the use of the absolute deviation around the median. Therefore, in this work the limit was set by the median absolute deviation (MAD_j) around the moving median (M_j) where j 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 Figure 6 shows the superior $Max_{threshold}$ and inferior $Min_{threshold}$ limits defined by (1)

$$
\begin{aligned}
Max_{threshold} &= M_j + N * MAD_j \\
Min_{threshold} &= M_j - N * MAD_j.
\end{aligned}
\tag{1}
$$

The length j of the window and the number of standard deviations denoted N were defined empirically, for each one of quantities analyzed (V, I, pf) as shown in Table 1. In the example (Figure 6), two samples were marked as outliers as they were not in between acceptance limits.

Time Series	j	N
V	96	1
pf	96	3
I	48	5

Table 1: Values of j and N for the removal of outliers using equation (1) defined empirically.

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 power distribution network reliability. For example, Figure 3 shows four MV feeders and three distribution relays. In this case if a fault occurs and FD02 looses its supply, the two normally open relays can be used to transfer its load to feeders FD01 and or FD03. However, for planning the expansion of capacity for the system, all data collected during the temporary load transfers must be discarded as they do not represent the normal operation in which the feeders are subjected and planned to work.

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. Figure 7 shows the three-phase current and power factor of one MV feeder and the voltage on the respective MV bus. In the example, a load transfer is conducted between July (dashed line) and the end of the period of study. The shift in characteristics for the current and power factor in July is very noticeable, whereas the voltage remains virtually unchanged. It is important to notice that this load transfer was conducted between different substation transformers.

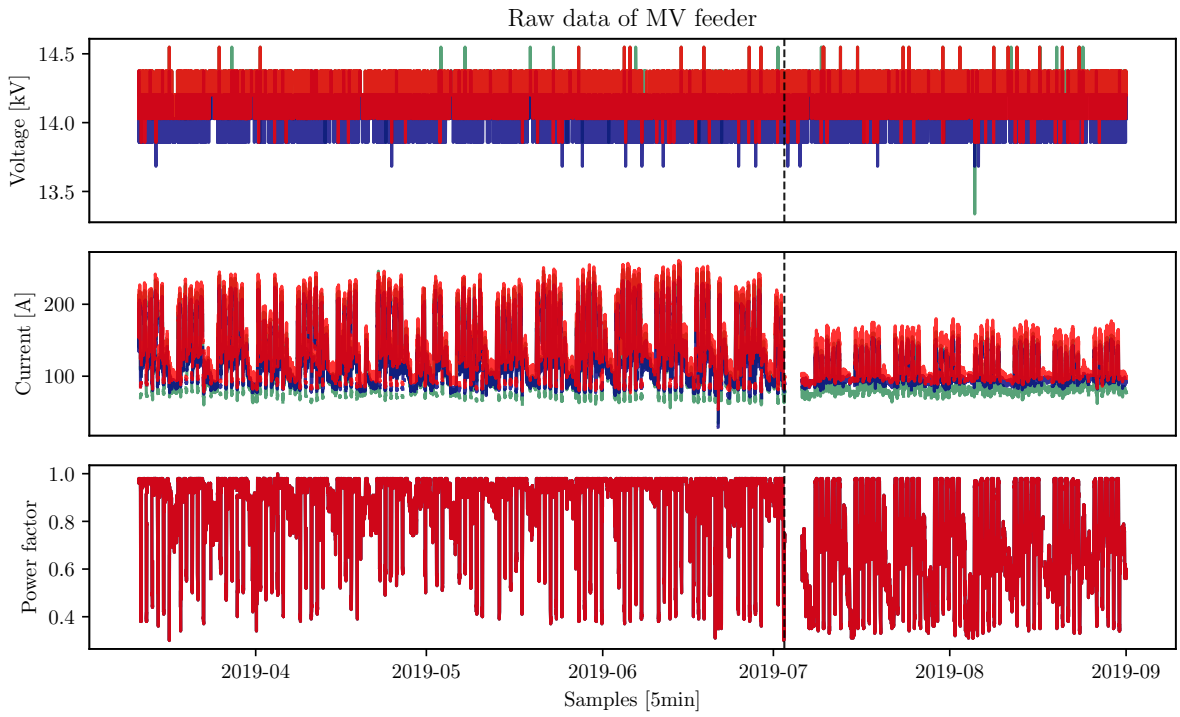

Figure 7: Example of raw data of MV feeder

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. For example, in Figure 3 the time series of the MV bus 1 was compared during normal operation of FD01 and FD02 and during the load transfer of FD01 or FD02. It is considered load transfers between feeders of same transformer and between different transformers (ex.: T2) or substations.

The procedure was done using a dependent sample t-test with a significance level (α) of 5% (Shier, 2004; Razavi-Far et al., 2020), 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 proposed imputation method has three main parts: initial processing and interpolation, data filling based on the ratio between phases, and data filling based on the 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, by the analysis of the raw data, it was assumed $N_{samples}$ = 18 (1.5 hours) as for this number of samples, the characteristics of the voltage, current, and power factor did not change dramatically (Peppanen et al., 2016). In a different dataset this number might have to be adjusted to fit the variance on the data being processed. A higher variance will indicate a smaller $N_{samples}$ and vice-versa. As will be 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). This first part, lays a solid foundation of consecutive filled samples for the next two steps of the imputation method.

After the first interpolation, the second stage uses the correlation between phases (ϕ_a, ϕ_b, ϕ_v) of the same quantity (V, I or 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.

The main idea is to use a period where all three-phases (ϕ_a, ϕ_b, ϕ_v) exist and calculate the proportion between them. Having the relationship between phases, if one or two are missing in a given timestamp i it is possible to use the remaining phase and the previous calculated ratio to fill the missing ones.

The number of samples used to calculate the ratio around the missing sample is an important parameter. For instance if a sample is missing in the afternoon it is best to use samples from that same day and afternoon to calculate the ratio and fill the missing sample. Unfortunately, there might be not enough samples in that period to calculate the ratio. Therefore, in this step, different periods T of analysis around the missing sample are considered: period of the day (dawn, morning, afternoon and night), month, and year.

The correlation between the feeder energy demand and the period of the day or the

season is very high. The increase in consumption in the morning and afternoon in industrial areas is expected as those are the periods where most factories are fully functioning. In residential areas, the consumption is expected to be higher in the evening; however, it is lower during the day's early hours. Furthermore, in the summer, a portion of the network (vacation destination) can be in higher demand. Nonetheless, in another period of the year (winter), the same area could have a lower energy demand. Therefore, if there is not enough information on that particular day to compute the ratio between phases, a good alternative is to use data from the month. Finally, given the amount of missing data for a particular feeder, the only option could be the use of the whole year to calculate the ratio between phases.

Regarding the minimum amount of data that a period should have to be valid it is assumed 50% for all three-phases (ϕ_a, ϕ_b, ϕ_v). The 50% limit of missing data for the period T is set to guaranty that there is more than half (majority) of valid samples, hence, it has enough data to estimate the ratio between phases with less probability of error. For example, the situation described were there was a missing sample in the afternoon, if there is less then 50% of valid samples on that particular afternoon the month will be used to calculate the ratio between phases. Additionally, if the month still does not have enough samples the year will be used and finally, as last resource, all the samples. In summary, if there is not enough data in part of the day where the missing sample is contained, the process is repeated for the month, the year, and then for the whole period of study.

This part of the algorithm will input all the missing samples where there is at least one adjacent sample and enough three-phase data around the missing sample to fill it 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. The T_i denotes the smallest valid period T around the sample i. Aditionally, $\overline{X_{\phi_a(T^i)}}$ denotes the average of samples of ϕ_a during a period T around i. Hence, $\frac{\overline{X_{\phi_a(T^i)}}}{\overline{X_{\phi_b(T^i)}}}$ is the ratio between ϕ_a and ϕ_b, and similarlly, $\frac{\overline{X_{\phi_a(T^i)}}}{\overline{X_{\phi_v(T^i)}}}$ is the ratio between ϕ_a and ϕ_v. The equation (2) shows that if for a timestamp i there is two adjacent samples, the third is calculated by the mean of the product of the adjacent samples and

its ratio regarding the missing one. The equation (3) uses the same idea, although in this case, as only one of the adjacent samples are *null* the result for $X^i_{\phi_a}$ will be using only one of the ratios. Hence, for a given timestamp i where $X^i_{\phi_a}$= *null* and $X^i_{\phi_b} \wedge X^i_{\phi_v} \neq null$

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

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

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

Consider Figure 8 that shows imputation process for the current time series with 15% of data loss. The first graph is the original data with no missing samples. The second one is the degraded version, and in this case, samples of ϕ_a are missing from the middle of April onwards. The next two graphs are the time series reconstructed by the proposed method and the Naive approach.

The example feeder shown in Figure 8 had the phases ϕ_b and ϕ_v with more load and therefore a higher current than ϕ_a. Hence, given that the other two-phases (ϕ_b, ϕ_v) were present during the period where ϕ_a were lost, the ratios between ϕ_a and ϕ_b, and between ϕ_a and ϕ_v are calculated. The period T used will vary for each sample, nevertheless, it is possible to infer that for the first missing samples the ratio in April will be used. For the other months the the period of study (January to July) was the one used to calculate the ratio. Hence, with proportion between phases calculated and using equations (2) and (3) samples of ϕ_a were filled. The result, as stated before, is shown on the Alg. curve in Figure 8. It is also important to notice that if all thee-phases were lost this procedure would not be possible.

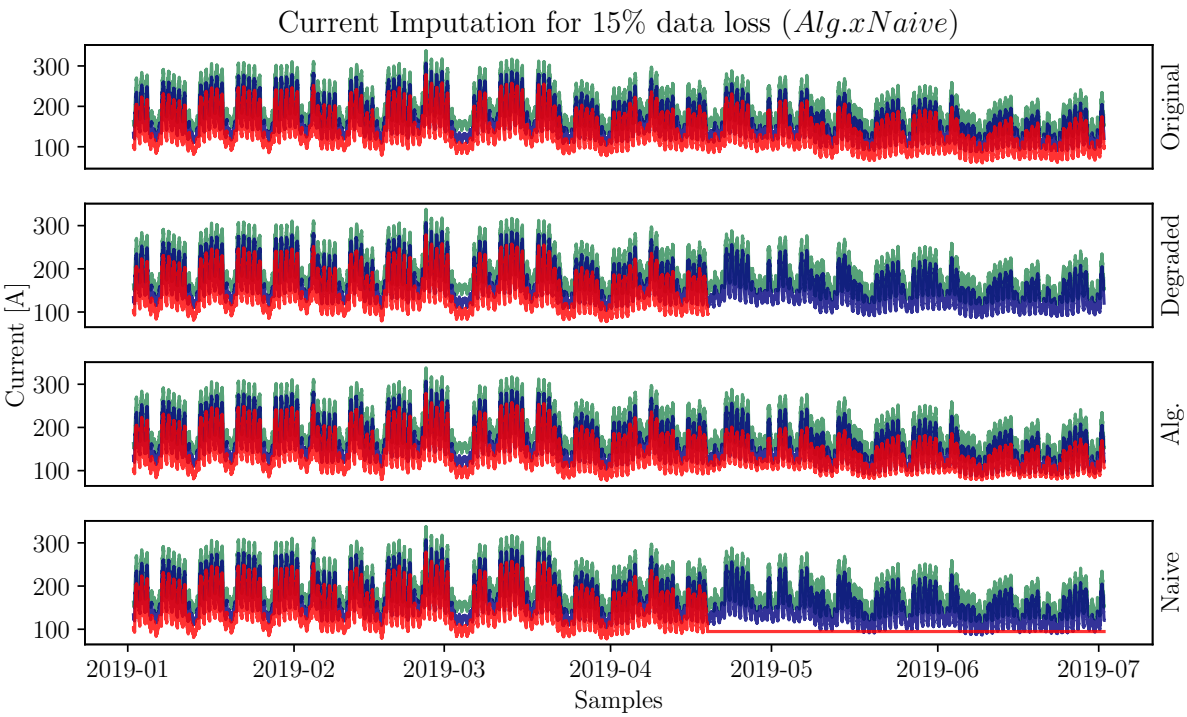

Figure 8: Example of a current time series with all the samples and its version degraded by 15%. It is also shown the proposed imputation method and the Naive results. The curves in red, green and blue, shows the phases ϕ_a, ϕ_b, and ϕ_v respectively.

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 Figure 11. In this part, a standard curve for each day of the week is calculated and normalized. This curve is the best approximation of the time series characteristics for each day. The standard curve will be used to fill the missing samples after being scaled by the maximum and minimum values of nearby days with no missing samples. The procedure and equations are described in the following paragraphs.

Equation (4) is used to calculate the normalized standard day of the week curve $S_{std}^{wd,\phi}$ where wd is the weekday, ϕ a specific phase and VD is the amount of valid days vd in the whole time series for a specific weekday. It is essential to notice that X_{ϕ}^{vd} 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 the number of days with no missing values in the time series of a feeder is less then three ($VD < 3$) for any day of the week (wd), this means that there is not enough data to calculate the $S_{std}^{wd,\phi}$. Therefore, an alternative is to find another feeder time series with similar characteristics

in the dataset and extract a standard curve of each day on the week.

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

Figure 9 shows the normalized standard weekday current curve for all the seven days of a given MV feeder.

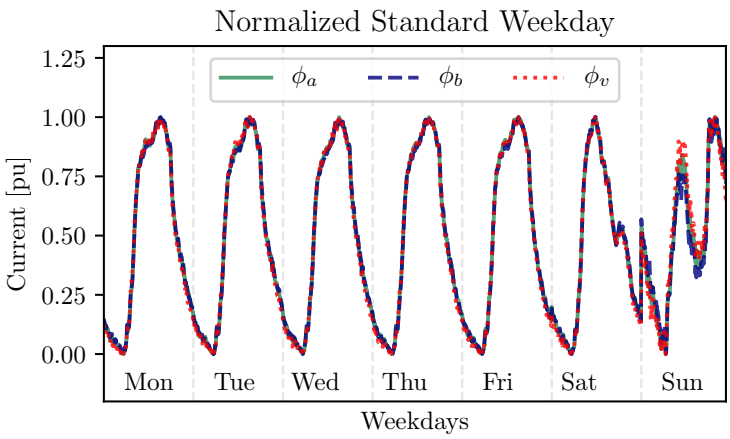

Figure 9: Example of seven $S_{std}^{wd,\phi}$ for current of a MV feeder

Another 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 minimum and maximum vectors (ζ and γ) are computed taking into account all the days in the time series. Additionally, in order to smooth any inconsistency, the moving average of two samples of the minimum and maximum vectors is used. Both pieces of information will be used to scale the $S_{std}^{wd,\phi}$ curve to input on a specific day, as shown in (5) and (6).

Equations (4), (5) and (6) are 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,

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

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

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

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).

Another situation is when consecutive samples in a giving day were not filled by previous steps (interpolation and proportion between phases) and the day has less then 50% of missing data. In this case, equations (4), (5) and (6) are used to fill a period of the day (dawn, morning, afternoon, or night). The difference is that the result of the equations (X_{ϕ}^{md}) is sliced in a particular period of interest that is missing of the day ($X_{\phi}^{md_{part}}$) and then it is used to fill the samples.

Figure 10 shows the graphical representation of the equations (4), (5) and (6). The first part shows that $S_{std}^{wd,\phi}$ of a given day is linearly scaled by the previous and next week minimum and maximum values of that same weekday. The second part shows the representation if the missing day is not between two valid days, hence, the closest one is used.

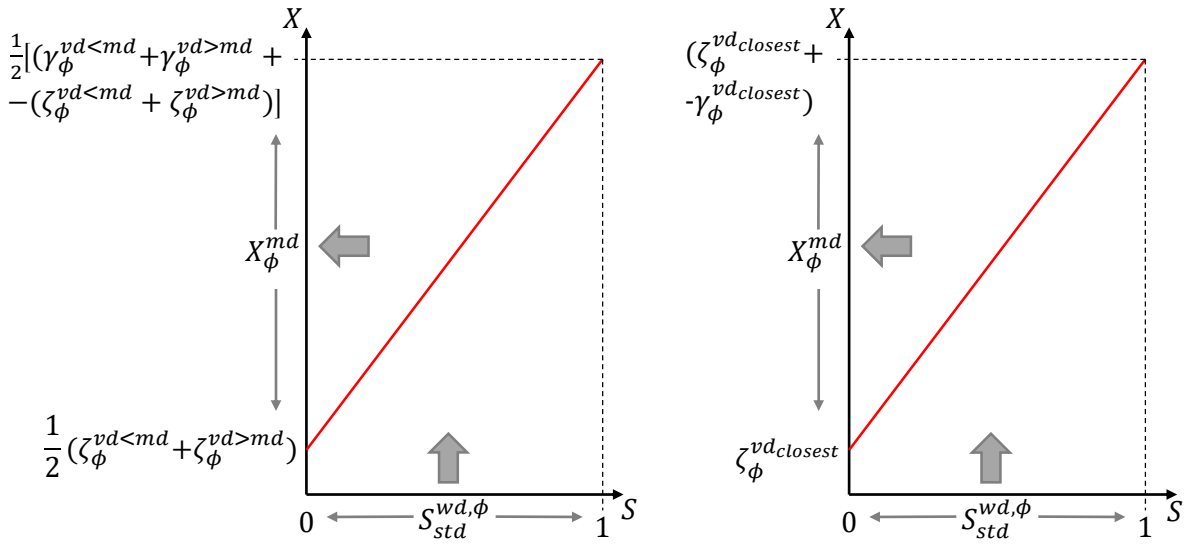

Figure 10: Graphical representation of equations 5 and 6.

The series of graphs shown in Figure 11 are the steps in the process of data imputation using the NSSC. Figure 11 a) shows 21 days of one-phase current time series (black) with three days of missing data. Figure 11 b) shows the maximum and minimum values of each complete day, vectors γ and ζ. Figure 11 c) shows the interpolation that will be used to scale the standard curves and fill days 9 (Tuesday), 13 (Saturday), and 14 (Sunday). Figures 11 d) and c) show in light grey the missing day curve obtained. Finally, Figure 11 c) shows the result of the method with no missing values in the current time series.

It is important to notice that, the proposed imputation method witch uses the NSSC 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. 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.

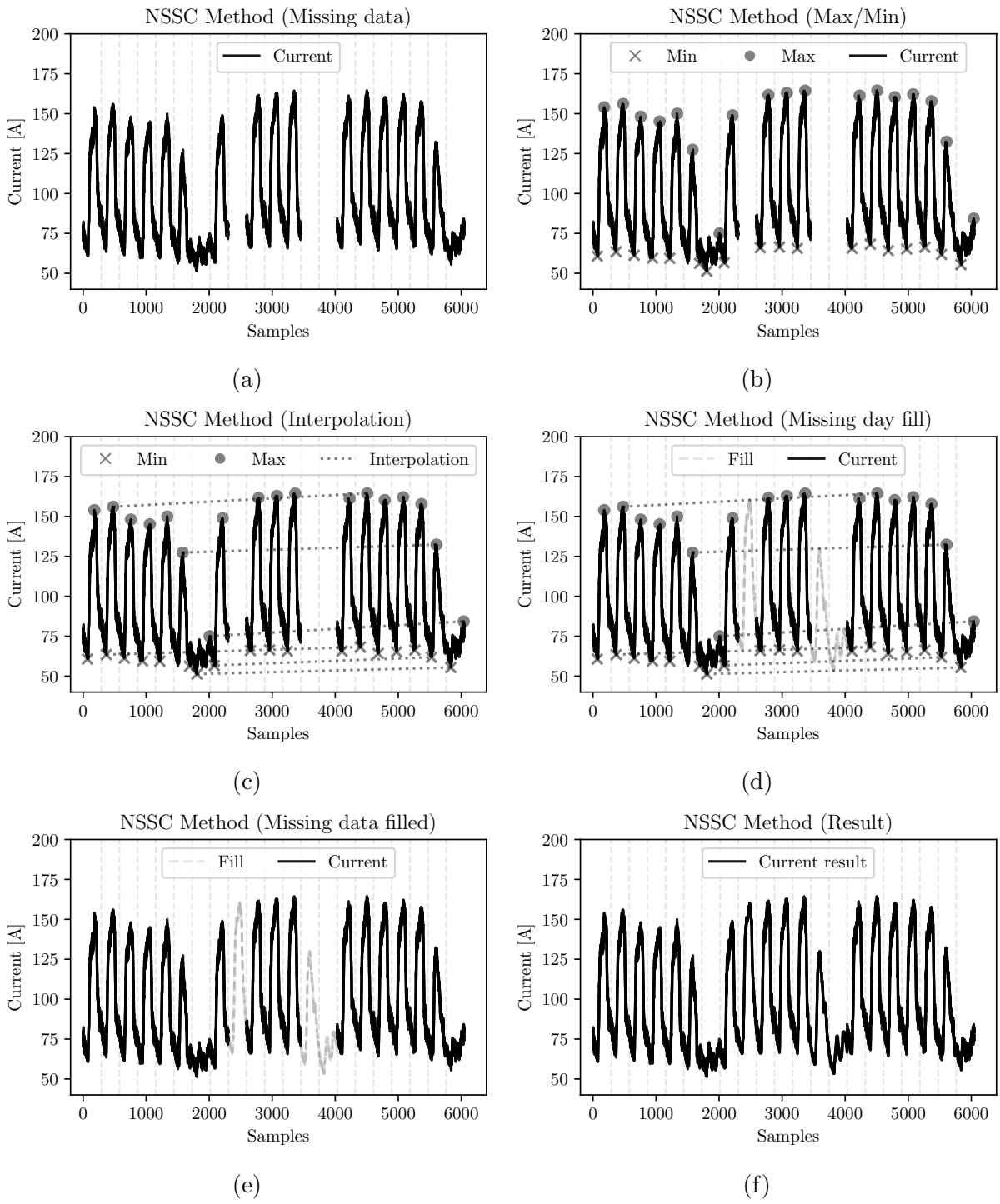

Figure 11: Example of using the normalized scaled standard weekday curve method, equations (4), (5), and (6), to fill missing days. The circles indicates the maximum values (γ) whereas the crosses indicates the minimum values (ζ) for each valid day (vd). Furthermore, the dark curve is the current time series and the light grey is the data inserted to fill missing values.

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. The summary of the imputation method presented in 2.4 is shown in algorithm 1.

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
    Synchronize time series
    for V, I and pf do
        Remove outliers
        Apply linear interpolation (N_samples = 18)
        for each phase (φ) do
            for each missing sample i (Applicable on any phase) do
                if X^i_{φ_a} = null and X^i_{φ_b} ∧ X_{φ^i_v} ≠ null then
                    Apply (2)
                if (X^i_{φ_b} ∨ X^i_{φ_v}) ≠ null then
                    Apply (3)
            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
            Compute 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 = ∞)
return
```

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 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 of consecutive samples. 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) as shown in section 3.2 (Murphy, 2012).

Knowing the PDFs, they can be used to tailor the imputation algorithm for optimal performance aiming for the types of missing sample with most probability and or that has most impact of the information loss. Additionally, it can be used to degraded a valid time series (a part or a whole time series with no missing samples) at different levels, as will be described in section 2.6, and test the imputation method comparing with the original data.

2.6 Imputation method test methodology

The missing data imputation method evaluation was conducted in a sub dataset for each quantity (V, I, pf), where there was no outlier or missing data. This subset is called original data and it is the time series known to have the true values. Figure 12 shows the flowchart for testing the imputation method. As described, the first step is to find, for each quantity, a portion of a feeder in the dataset with no inconsistencies. These parts

will compose the original subset. Afterwards, this subset is degraded in different levels by the PDF of missing data extracted from all the dataset, as discussed in section 2.5. With the degraded time series, the imputation method proposed in section 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^2, MAPE, and RMSE (Lai et al., 2019; Jadhav et al., 2019; Razavi-Far et al., 2020; Mccoy et al., 2018). Finally, it is important to mention that the time series were degraded from the following levels of data loss: $1\%, 2\%, 3\%, 4\%, 5\%, 10\%, 15\%, ..., 85\%, 90\%, 95\%$. It is important to mention that the 100% data loss level was not tested as it is assumed that some information of the original time series is required to input data on the missing samples.

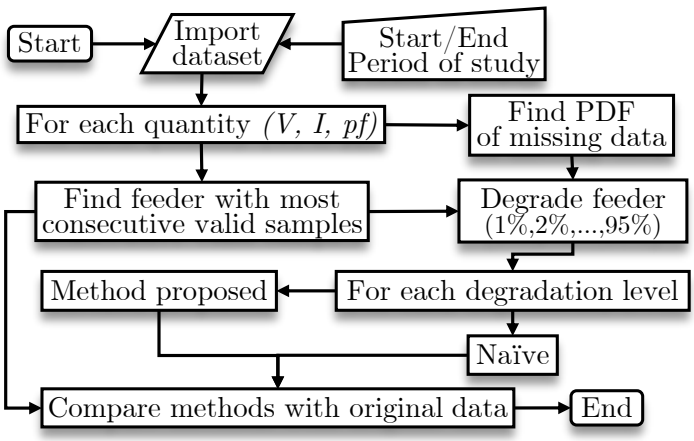

Figure 12: Flowchart of the imputation test methodology

3 Results and Discussion

In this chapter, the results obtained in this work are described. First, is presented the analysis of the impact of load transfers on the substation bus voltage. Furthermore, the conclusions and insights obtained by the study of missing data in the dataset regarding the number of phases, length, and the percentage of data loss per feeder are discussed. Finally, as described in chapter 2, the imputation method results are shown, including the comparison with the Naive approach.

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 substation three-phase voltage during normal operation and load transfer of any of the related medium-voltage feeders.

The result obtained confirms the premises by which the DSO operates, where the load transfer would be performed only when it is possible to maintain quality supply standards for all the feeders involved. Therefore, no over or under voltages would be noticed by the customers. The idea is that it is would not be preferable to degrade all customer's supply of a particular substation in order to receive load from another substation. Another important consideration is the use Load Tap Changer between the substation distribution transformer and the MV Bus. The correction of voltage variations is done automatically to accommodate differences in load. Hence, the voltage regulation makes transferring load possible without jeopardizing the quality supply for the customers involved as long as the amount of load shifted is maintained within the capabilities of the substation.

3.2 Analysis of missing data

The analysis of missing data in the dataset of 459 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, loses 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 13 shows the percentage of occurrences in the dataset of each type of data lost, whereas Figure 14 shows the percentage of occurrences of each length of consecutive data sample 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 Figure 13. The one and two-phases sample loss are the ones that can be filled using the proportion between phases.

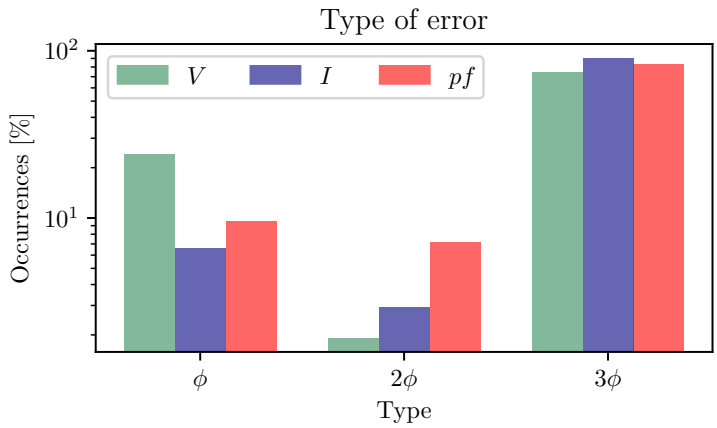

Figure 13: Percentage of occurrences of one, two and three-phase data loss.

The majority of three-phase data loss shown in Figure 13 can be explained by the stop of the relays for maintenance or a communication interruption that prevents all the three phases to be acquired by the DSO. Additionally, the load transfer event will also contribute for this result given that it affects all phases.

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 minutes. 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 3×10^4 samples is sufficient to compromise the analysis of the feeder with months of missing values.

For the current and power factor time series, there is a more extensive occurrence of multiple consecutive missing samples. That is explained as the load transfers were not removed from the voltage time series, and these events can range from a few minutes up to months depending on the severity of the incident.

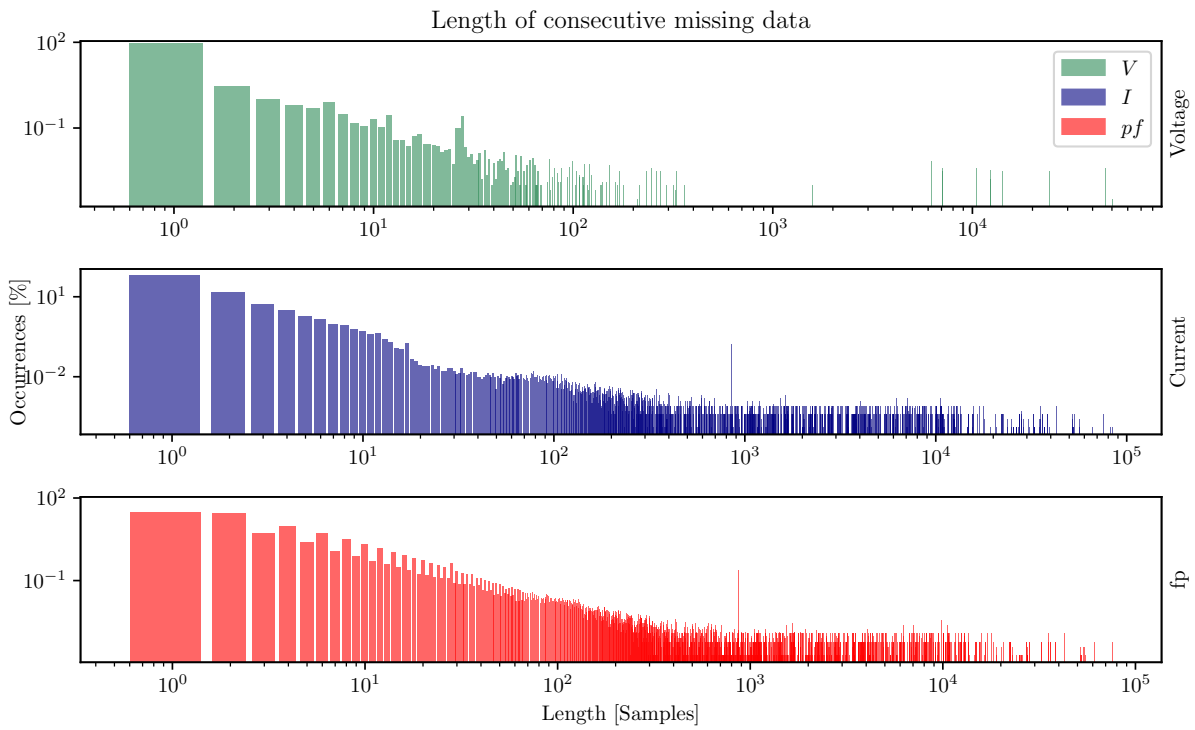

Figure 14: Percentage of occurrences of each consecutive missing samples length.

Figure 15 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. This analysis indicates that an algorithm that can be effective in treating up to 60% of data degradation in a feeder will contemplate most cases in the database studied.

Similarly, as explained for the consecutive missing data length, the voltage time series lost less data than the current and power factor given that the load transfers were not removed from its time series.

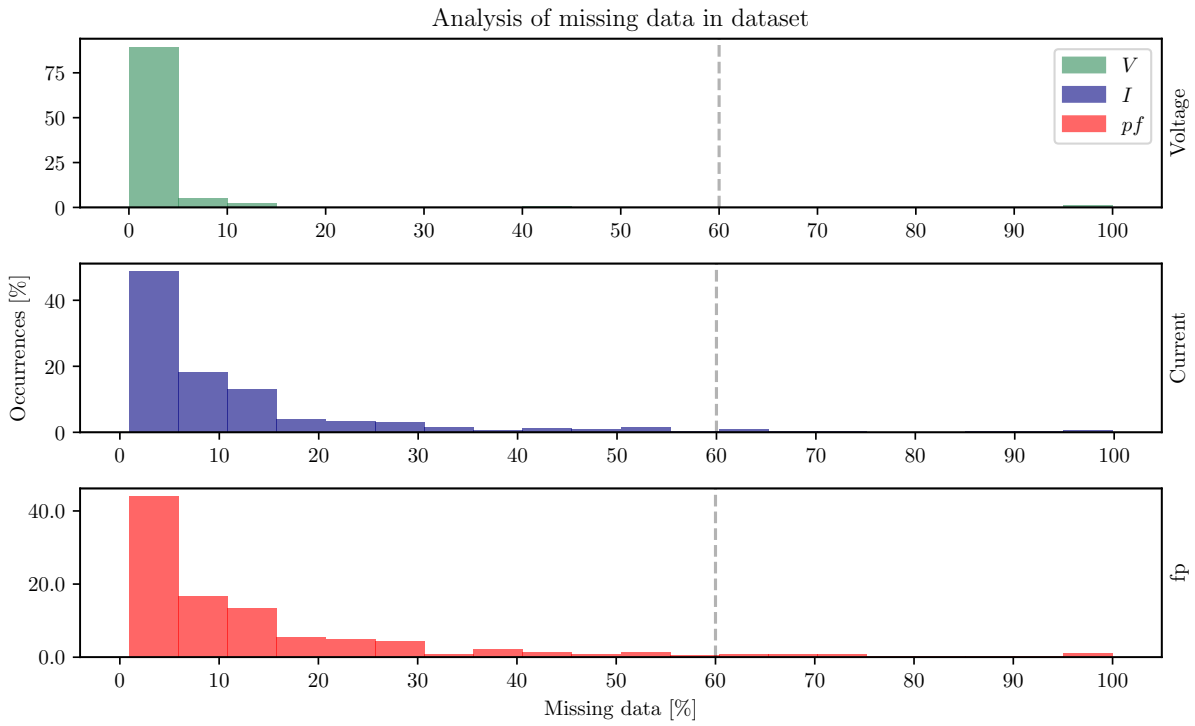

Figure 15: Percentage of missing values per feeder in the dataset.

3.3 Imputation method

The results shown in this section were obtained after 20 executions of the method on each degradation level, as described in Figure 12. Figures 16, 17, and 18 show the average result whereas Figure 19 shows the statistical analysis of multiple executions. It is important to notice that, as the degradation level is obtained randomly using the Probability Density Functions (PDFs), multiple executions would assure that the result is converging to the actual performance of the algorithm. Additionally, it is important to notice that the feeder chosen to perform the test was not the same for each electrical quantity. For the voltage time series the feeder had consecutive non missing samples from May to September, whereas for current, a different feeder was chosen with six months of valid data ranging from January to July. Finally, for the power factor a five months period was studied from February until July which was also from another feeder in the dataset.

The use of different feeders was an alternative to have the largest period of consecutive data available for testing on each individual electrical quantity. Figures 16, 17 and 18 show the result for the proposed method in section 2.4 compared with the Naive approach using real feeder data and tested as described in section 2.6.

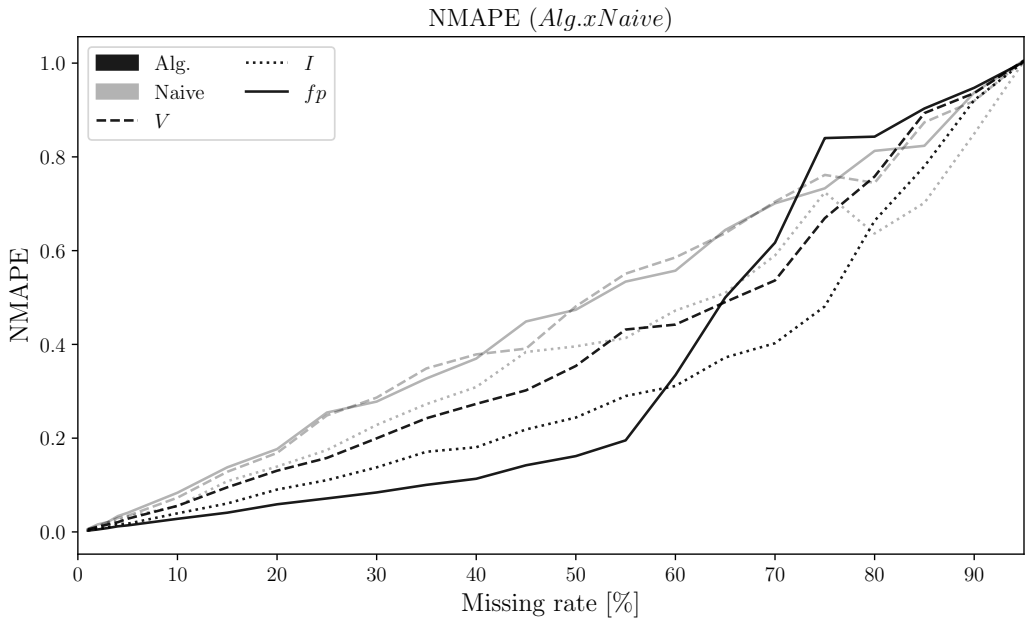

Figure 16: Evaluation of the method proposed for different degradation levels using MAPE.

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.

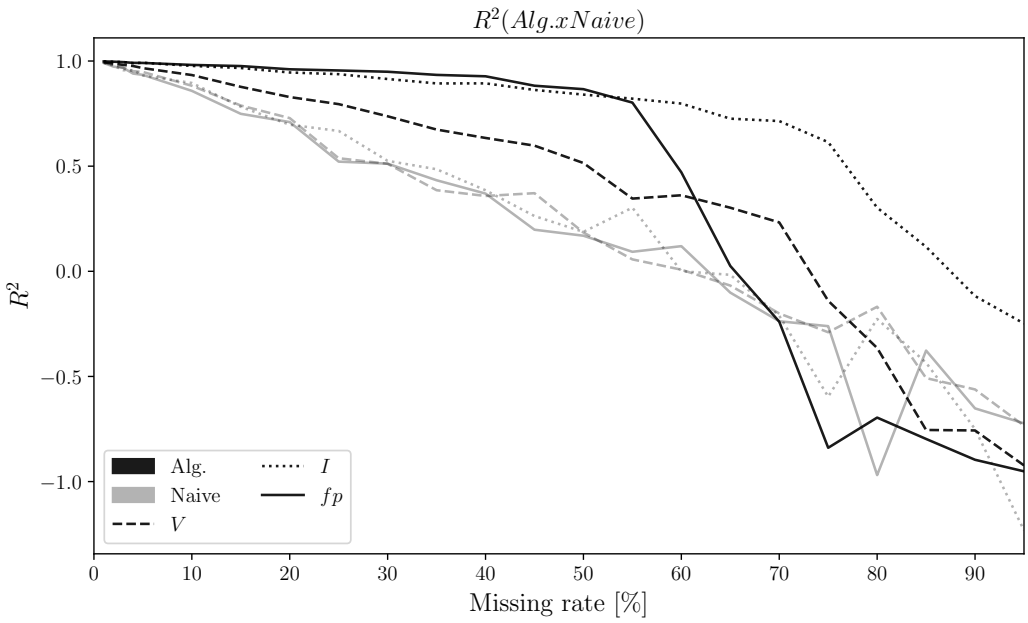

Figure 17: Evaluation of the method proposed for different degradation levels using R^2.

It is essential to notice that it is expected for small degradation levels (up to 5%) that the two methods would have similar low values of RMSE and MAPE. Additionally, in the first part of Figure 17, which represents the minimum amount of missing values, it is expected that R^2 would be close to one. On the other hand, for massive degradation levels (larger than 85%), very few methods would likely fill all the missing samples. Regarding the dataset studied, the loss of 85% would comprehend losing ten months in the year, and with the other two months be able to restore the characteristics of the whole period of study.

Furthermore, regarding the qualitative analysis of Figures 16 and 18, a good result would be for a method to keep its performance curve for most of the degradation levels close to one for R^2 and zero for the normalized RMSE and MAPE. Therefore, decreasing its performance, although rapidly, only for high degradation levels.

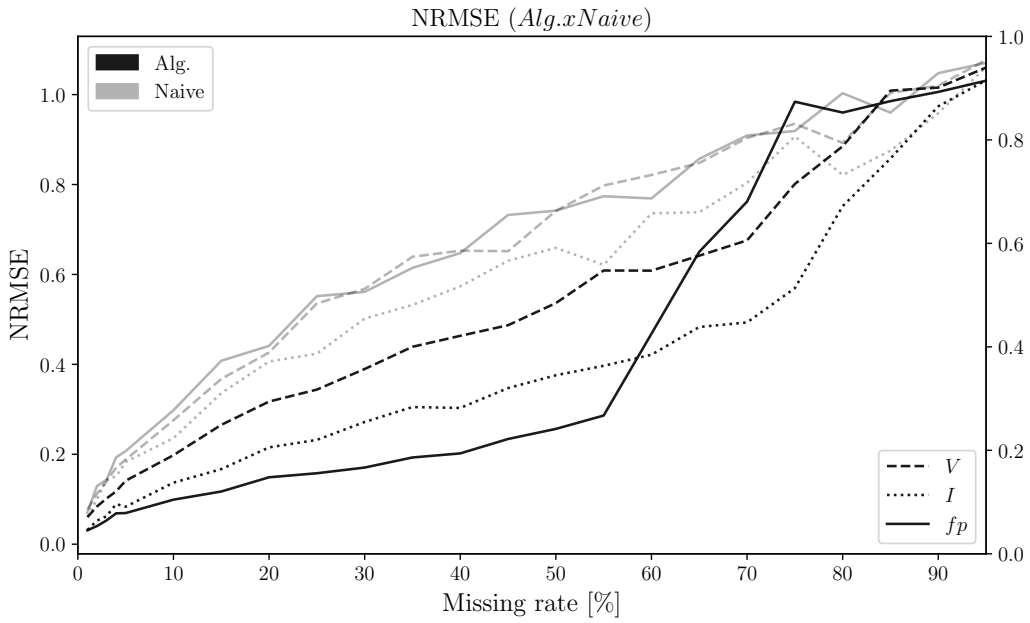

Figure 18: Evaluation of the method proposed for different degradation levels using RMSE.

Figure 19 shows the statistical differences between the proposed method and the Naive approach using a dependent sample t-test with a significance level (α) of 5%. For the power factor, the point of no statistical difference between the methods starts 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 (Jadhav et al., 2019; Razavi-Far et al., 2020; Chang et al., 2015).

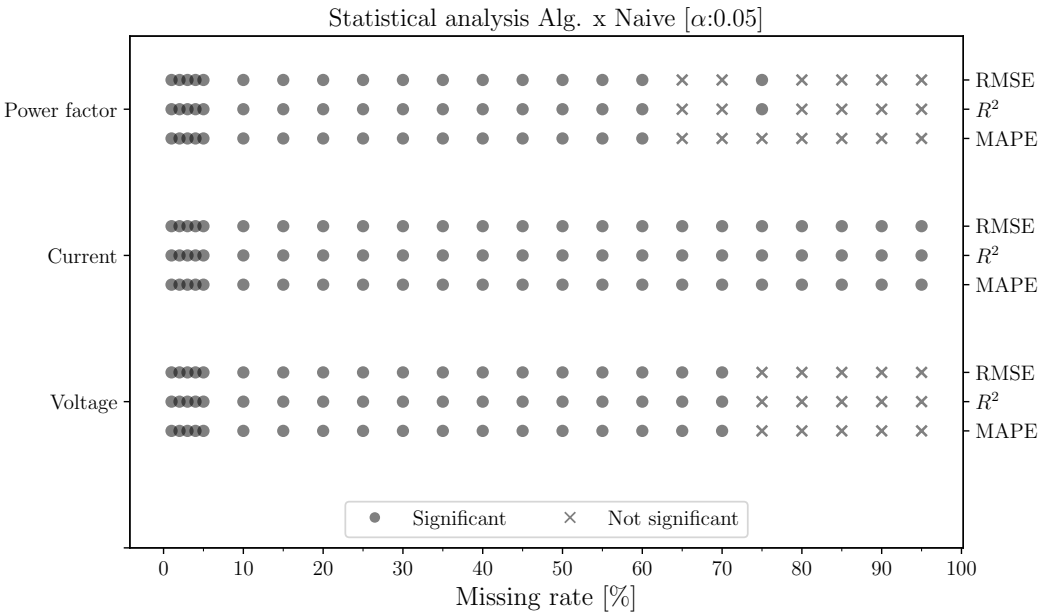

Figure 19: Statistical analysis: performance of the algorithm proposed and the Naive approach.

Furthermore, although it is expected that for extensive periods of consecutive missing samples, the Naive approach would not be a suitable imputation method, for small lengths, as shown in section 1.3 it is a method commonly used. Additionally, section 3.2 demonstrates that most missing samples are of short periods, therefore it is still valid to compare the two methods for the degradation levels and characteristics of missing data studied in this work. Nevertheless the imputation method proposed had better performance not only in larger amounts of degradation levels but also for the small ones as shown in Figures 16 to 18.

Figure 20 shows six months of a three-phase current time series. The first part shows the original data. The second part shows the 25% degraded version with long periods of three-phase and one-phase data loss. The degraded version of the original data was obtained by the method discussed in 2.6. The last two graphs in Figure 20 shows the results of the proposed method in 2.4 and the Naive approach of missing data imputation.

The periods around February and July lost all three-phases of the current data. Therefore, for these periods, the alternative is to apply the NSSC imputation method. For the periods around March and from April to June there is still one phase remaining, hence, the proportions between phases is used to fill those samples. For short periods of missing

values ($N_{samples} = 18$), although not clearly shown, the linear interpolation is used.

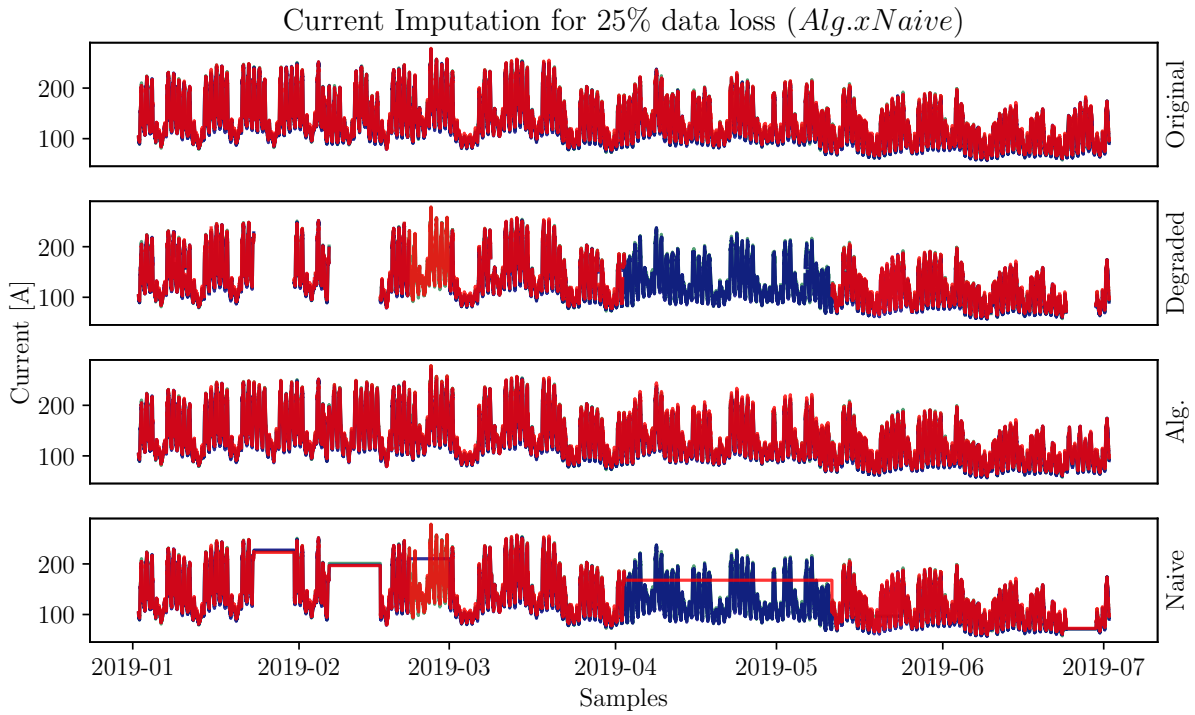

Figure 20: Example of an current time series with all the samples and its version degraded by 25%. It is also shown the proposed imputation method and the Naive results. The curves in red, green and blue, shows the phases ϕ_a, ϕ_b, and ϕ_v respectively. For this particular case the method proposed had R^2: 0.937, MAPE: 0.110, and RMSE: 0.231. On the other hand, the Naive method had R^2: 0.667, MAPE: 0.174, and RMSE: 0.423.

In Appendix A, the comparison of the original data, the degraded version by the PDFs of missing data, and the imputed time series with the algorithm proposed and the Naive approach are shown. Figures 22 to 90 of the Appendix shows every degradation level tested in one execution of the proposed method for V, I and pf.

4 Conclusion and future work

In this study, a method of data preprocessing and missing sample imputation for medium voltage feeders is proposed. The method was based on the analysis of 459 MV feeders of a utility company in Brazil, and considered the impact of missing values after the removal of outlier and load transfer. It was verified that the information of the medium voltage feeders sampled at the 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 to perform operations with the voltage, current, and power factor of each feeder. Additionally, 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. In the second part, samples that do not respect the physical and or theoretical constraints of the system are removed. In the third part, a statistical method is applied 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. Furthermore, the linear interpolation was used for the samples that could not be filled by the ration between phases. For more extensive periods of three-phase data loss, the missing values were imputed by the normalized scaled standard weekday curve (NSSC). This last part takes advantage of the correlation between the quantities and the weekdays. Furthermore, the proposed three-part method was compared with the Naive approach and showed promising results. For voltage and power factor, the results were statistically significant for up to 60% of feeder degradation, whereas, for current, the method is statistically better for up to 95%.

As a proposition of future work, measuring the similarity between feeders using cus-

tomer's characteristics, geographical area, length, and any external information besides the actual electrical data would be extremely valuable. The similarity measurement would provide the ability to find a compatible feeder to extract the standard weekday curve and input data for feeders with too many missing samples. Additionally, other methods of outlier removal could be aggregated to those existing to improve the results obtained.

References

Mehran Amiri and Richard Jensen. Missing data imputation using fuzzy-rough methods. *Neurocomputing*, 205:152–164, 2016. doi: 10.1016/j.neucom.2016.04.015.

Aneel. Prodist, Jan 2017. URL `https://www.aneel.gov.br/prodist`.

Faraj Bashir and Hua-Liang Wei. Handling missing data in multivariate time series using a vector autoregressive model based imputation (var-im) algorithm: Part i: Var-im algorithm versus traditional methods. pages 611–616, 06 2016. doi: 10.1109/MED.2016.7535976.

Chia-Yang Chang, Cheng-Ru Wang, and Shie-Jue Lee. Novel imputation for time series data. *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2015. doi: 10.1109/icmlc.2015.7340675.

Grigoras Gheorghe, G. Cartina, Elena Bobric, and C. Barbulescu. Missing data treatment of the load profiles in distribution networks. *2009 IEEE Bucharest PowerTech: Innovative Ideas Toward the Electrical Grid of the Future*, 06 2009. doi: 10.1109/PTC.2009.5282021.

Turan Gonen. *Electric Power Distribution System Engineering*. CRC Press, 2007.

LLC Google. A history of machine learning, 2020. URL `https://cloud.withgoogle.com/build/data-analytics/explore-history-machine-learning/`.

Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan Kaufmann, 2012.

Nathalie Huyghues-Beaufond, Simon Tindemans, Paola Falugi, Mingyang Sun, and Goran Strbac. Robust and automatic data cleansing method for short-term load forecasting of distribution feeders. *Applied Energy*, 261:114405, 2020. doi: 10.1016/j.apenergy.2019.114405.

Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33 (10):913–933, 2019. ISSN 10876545. doi: 10.1080/08839514.2019.1637138. URL `https://doi.org/10.1080/08839514.2019.1637138`.

Nicolas Kong, Maxence Bocquel, Thibaut Barbier, Robin Girard, Elena Magliaro, Georges Kariniotakis, Guillaume Pelton, and Pierre Cauchois. Long-term forecast of local electrical demand and evaluation of future impacts on the electricity distribution network. *CIRED - Open Access Proceedings Journal*, 2017(1):2401–2405, 2017. doi: 10.1049/oap-cired.2017.0743.

Xiaochen Lai, Xia Wu, Liyong Zhang, Wei Lu, and Chongquan Zhong. Imputations of missing values using a tracking-removed autoencoder trained with incomplete data. *Neurocomputing*, 366, 07 2019. doi: 10.1016/j.neucom.2019.07.066.

Christophe LEYS, Olivier KLEIN, Philippe BERNARD, and Laurent. LICATA. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation a round the median. *Journal of Experimental Social Psychology*, pages 764–766, 2019.

John T. Mccoy, Steve Kroon, and Lidia Auret. Variational autoencoders for missing data imputation with application to a simulated milling circuit. *IFAC-PapersOnLine*, 51 (21):141–146, 2018. doi: 10.1016/j.ifacol.2018.09.406.

Scott Miller and Donald Childers. *Probability and random processes: with applications to signal processing and communications*. Elsevier Academic Press, 2004.

Gregorio Muñoz-Delgado, J. Contreras, and José Arroyo. *Distribution System Expansion Planning*, pages 1–39. 04 2018. ISBN 978-981-10-7055-6.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. The MIT Press, 1 edition, 2012. ISBN 0262018020,9780262018029.

Jouni Peppanen, Xiaochen Zhang, Santiago Grijalva, and Matthew J. Reno. Handling bad or missing smart meter data through advanced data imputation. *2016 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2016*, pages 0–4, 2016. doi: 10.1109/ISGT.2016.7781213.

Roozbeh Razavi-Far, Boyuan Cheng, Mehrdad Saif, and Majid Ahmadi. Similarity-learning information-fusion schemes for missing data imputation. *Knowledge-Based Systems*, 187:104805, 2020. doi: 10.1016/j.knosys.2019.06.013.

Jeanne Saunders, Nancy Morrow-Howell, Edward Spitznagel, Peter Dore, Enola Proctor, and Richard Pescarino. Imputing missing data: A comparison of methods for social work researchers. *Social Work Research*, 30:19–31, 03 2006a. doi: 10.1093/swr/30.1.19.

Jeanne A. Saunders, Nancy Morrow-Howell, Edward Spitznagel, Peter Doré, Enola K. Proctor, and Richard Pescarino. Imputing missing data: A comparison of methods for social work researchers. *Social Work Research*, 30(1):19–31, 2006b. ISSN 10705309. doi: 10.1093/swr/30.1.19.

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

Eduardo Lenhart Vargas. 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, 2015.

Wei-Tzer Huang Wen-Chih Yang. An enhanced load transfer scheme for power distribution systems connected with distributed generation sources. *WSEAS TRANSACTIONS on CIRCUITS and SYSTEMS*, 2011.

Tieyan Zhang, Li Liu, Huaguang Zhang, and Yuan Zhang. Incomplete load data processing in distribution system. *2006 IEEE International Conference on Networking, Sensing and Control*. doi: 10.1109/icnsc.2006.1673147.

Appendices

Apendix A

The results obtained from a single execution of the test procedures are shown in this section. The results shown in section 3.3 were obtains after twenty executions, nevertheless, this figures give a glimpse on the level of degradation on each time series studied (V, I, pf). The data shown in this section is presented as demonstrated by Figure 21, where the first graph shows the original data, the second, the same time series degraded by a level of loss, and the following two graphs shows the NSSC and the Naive method, respectively.

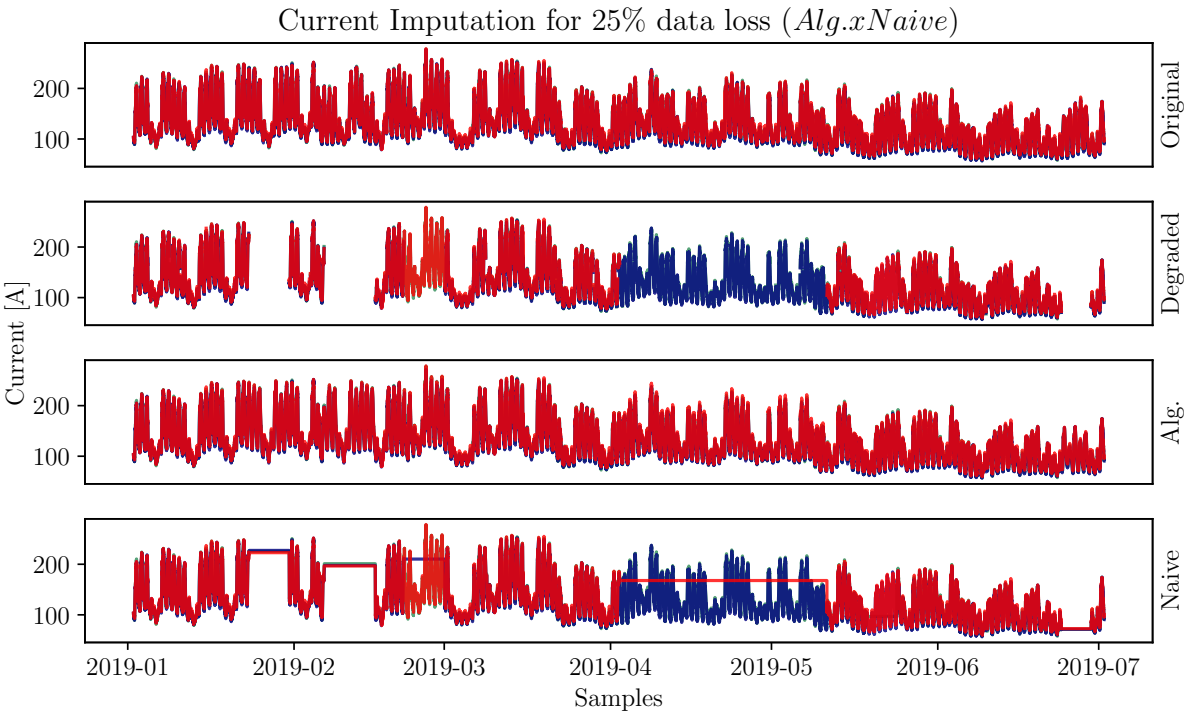

Figure 21: Example of the structure of data shown in the figures of Appendix A.

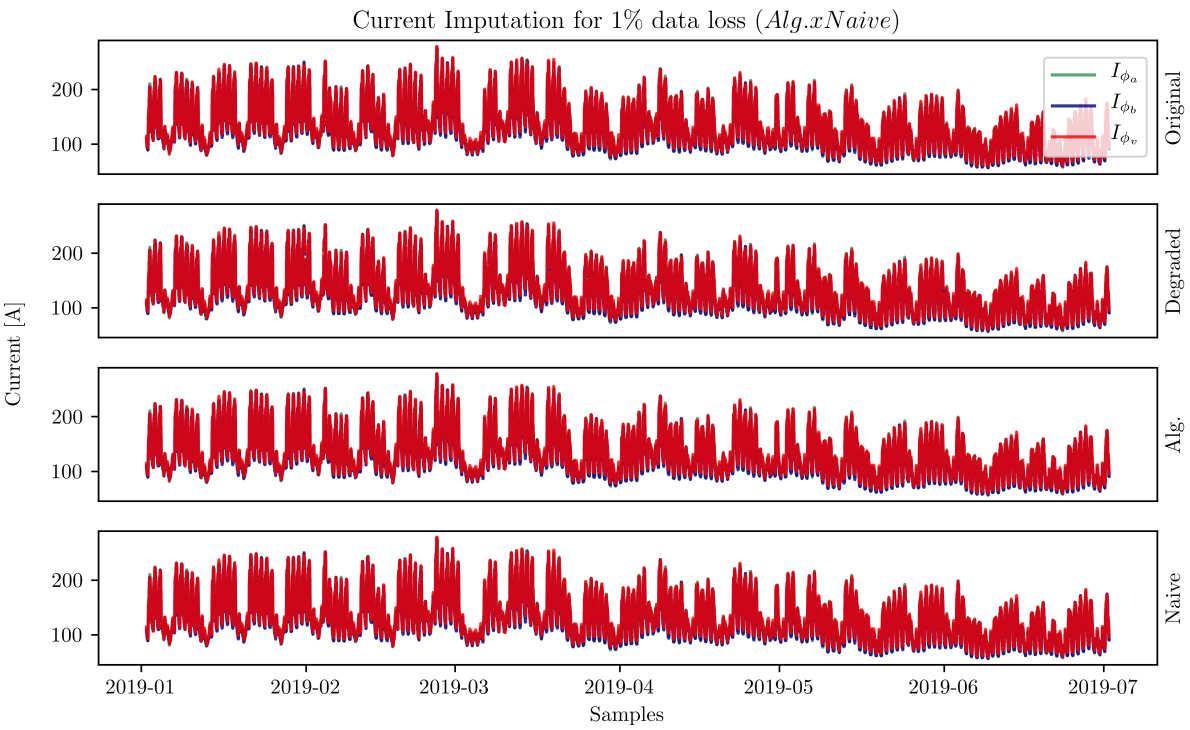

Figure 22: Example of 1% data loss on the current time series.

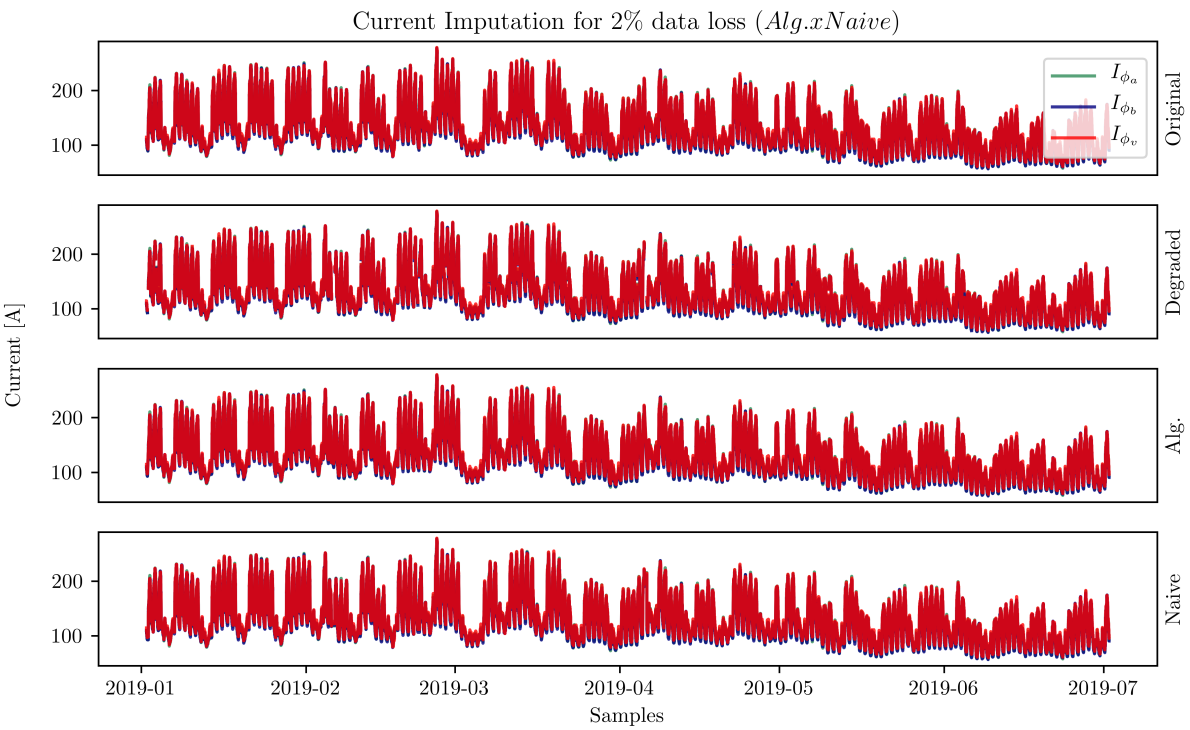

Figure 23: Example of 2% data loss on the current time series.

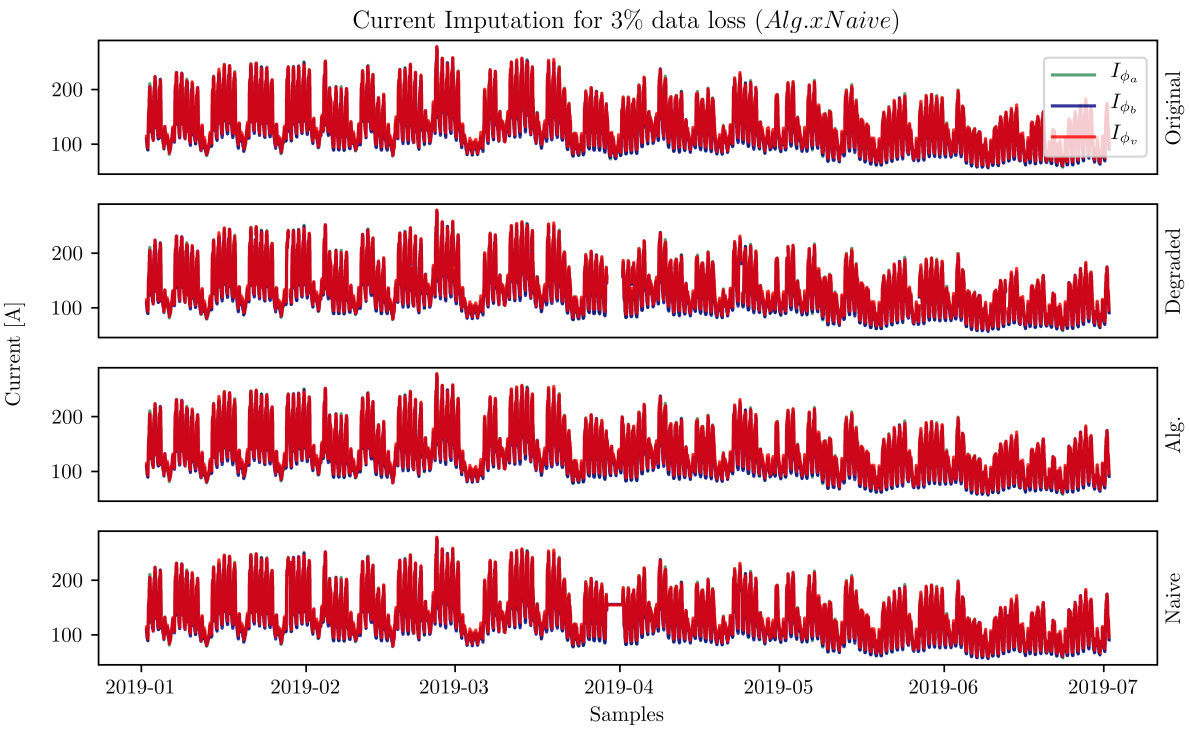

Figure 24: Example of 3% data loss on the current time series.

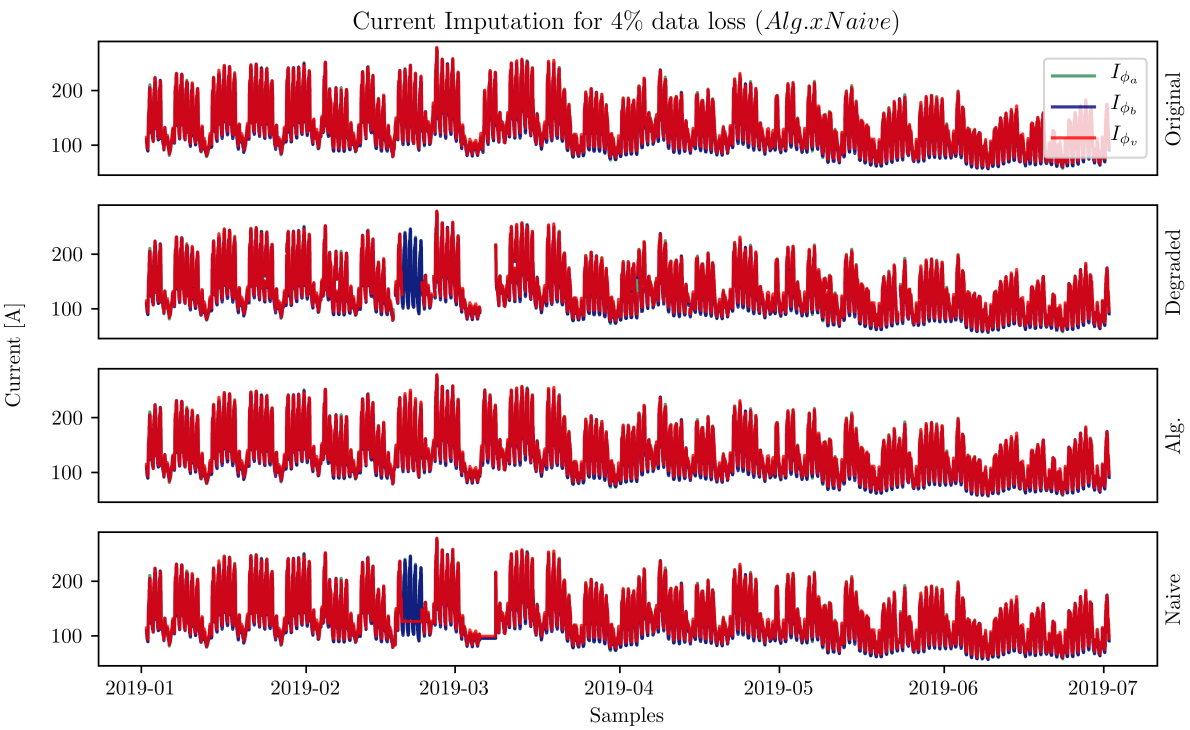

Figure 25: Example of 4% data loss on the current time series.

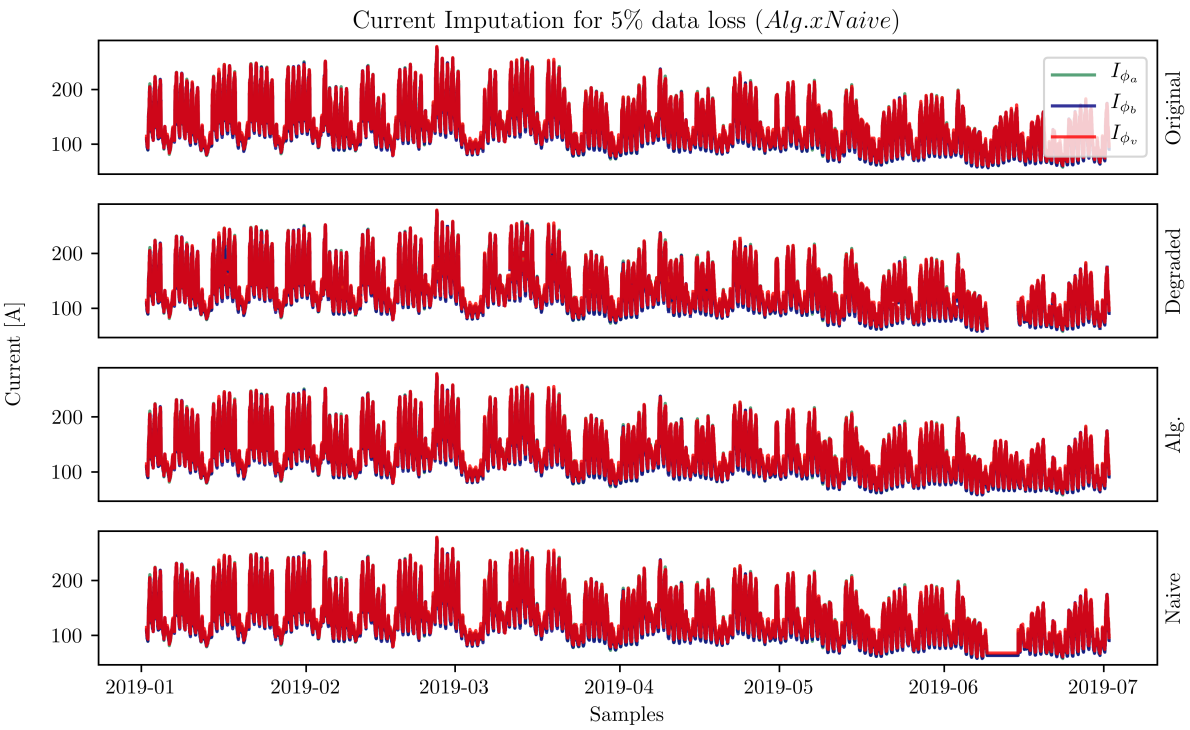

Figure 26: Example of 5% data loss on the current time series.

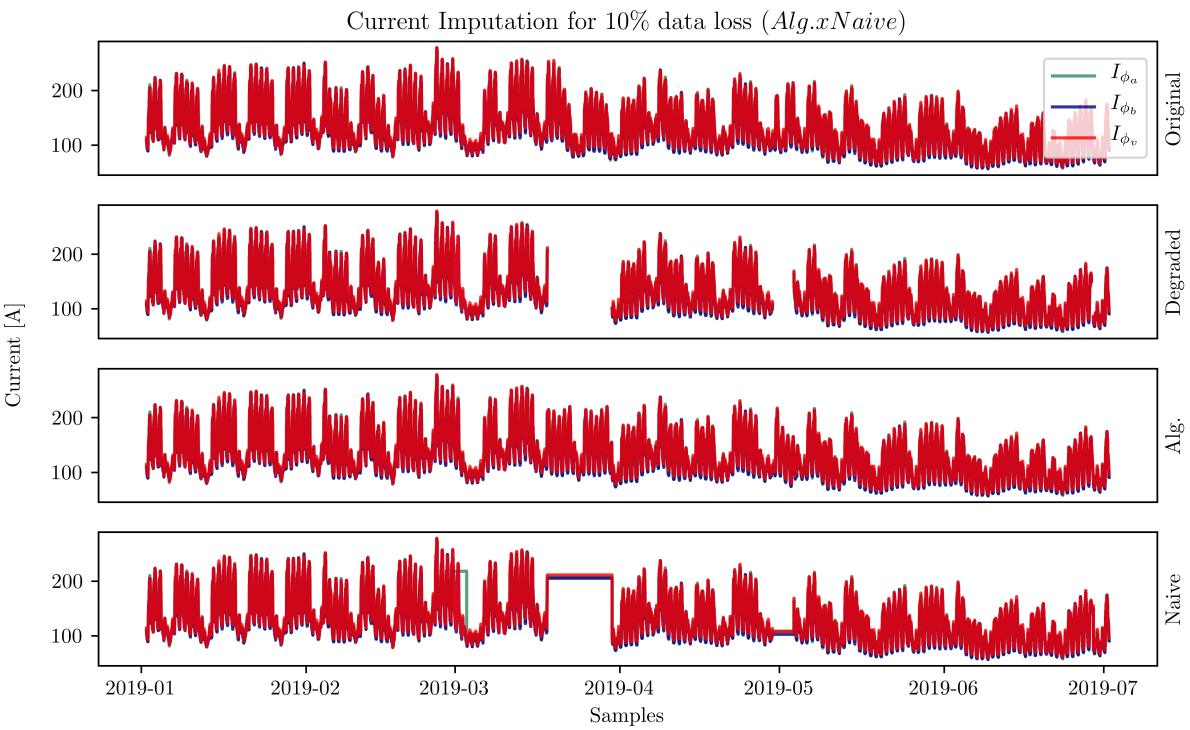

Figure 27: Example of 10% data loss on the current time series.

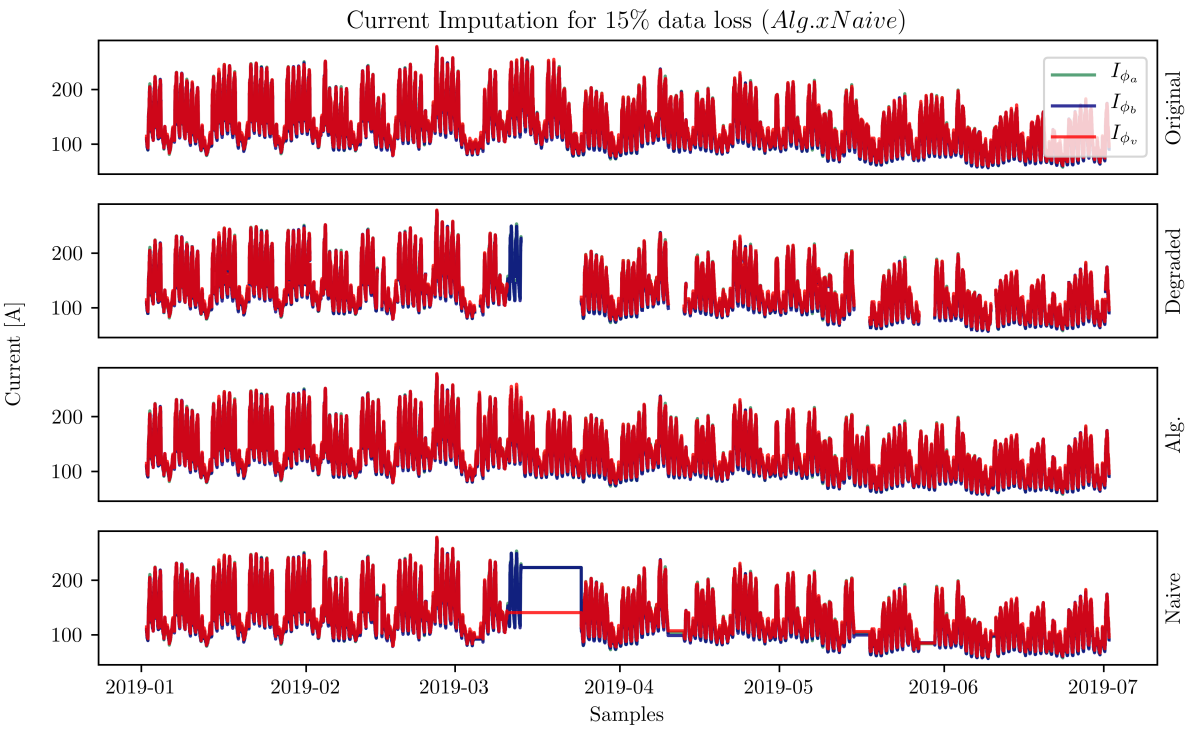

Figure 28: Example of 15% data loss on the current time series.

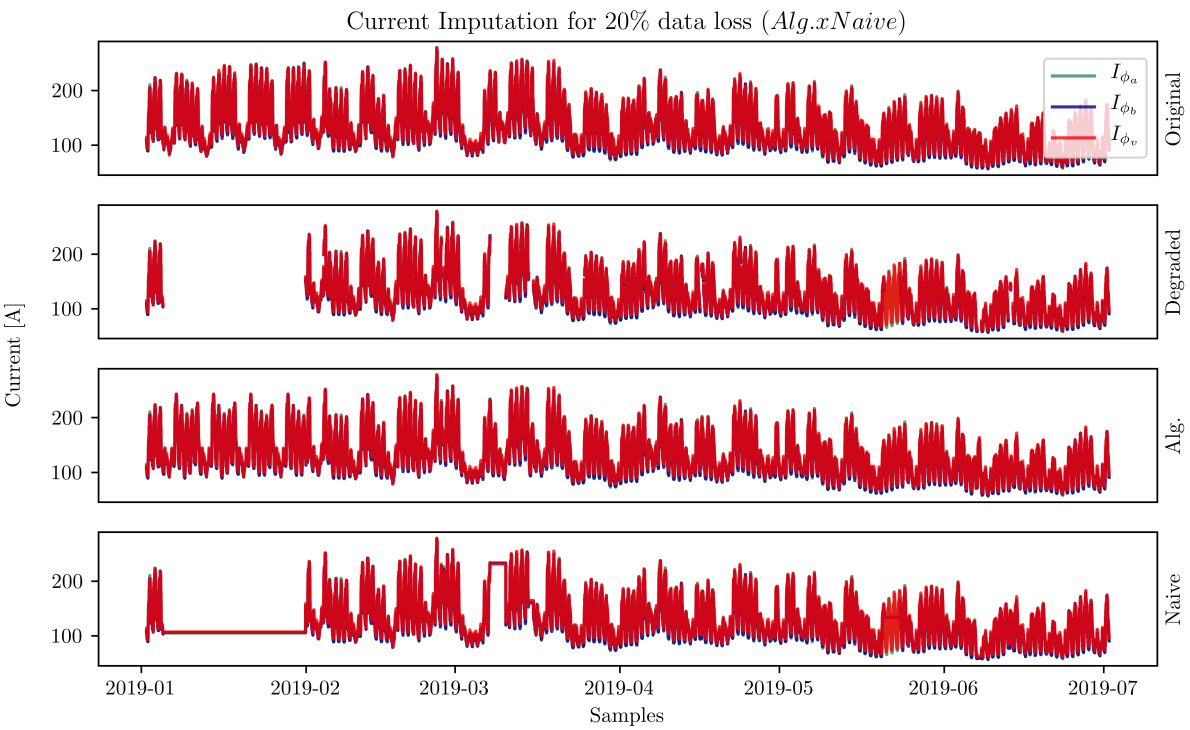

Figure 29: Example of 20% data loss on the current time series.

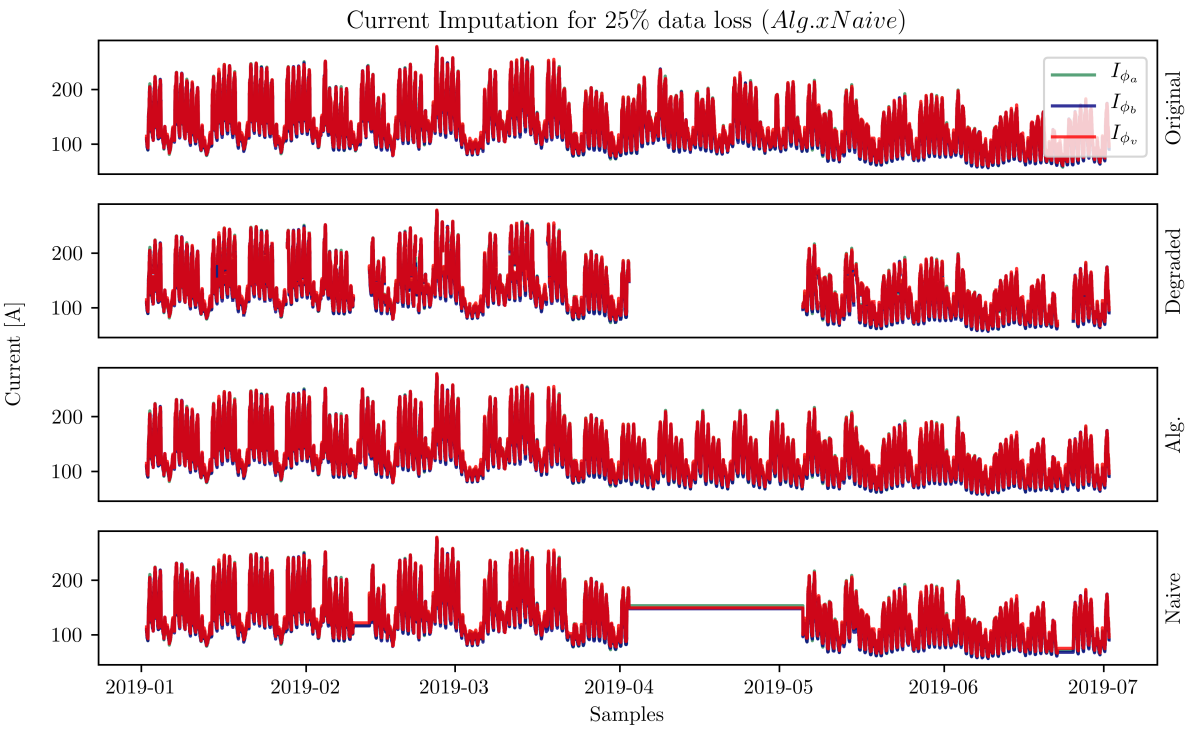

Figure 30: Example of 25% data loss on the current time series.

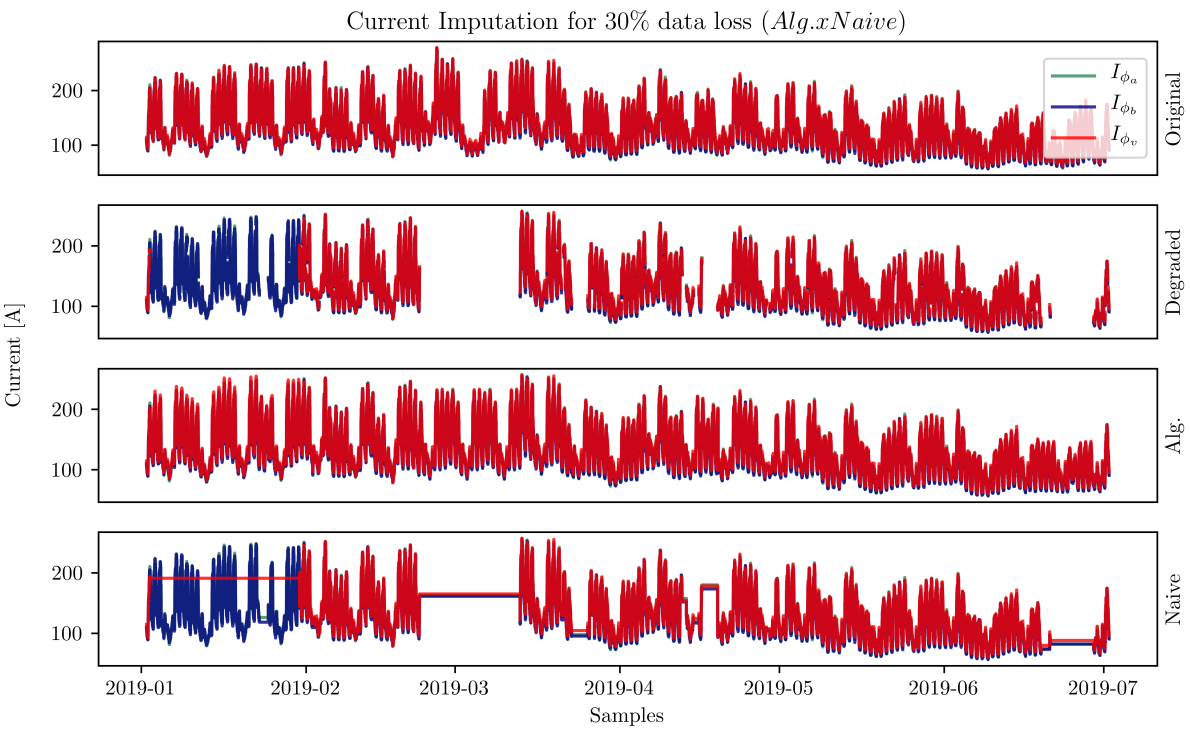

Figure 31: Example of 30% data loss on the current time series.

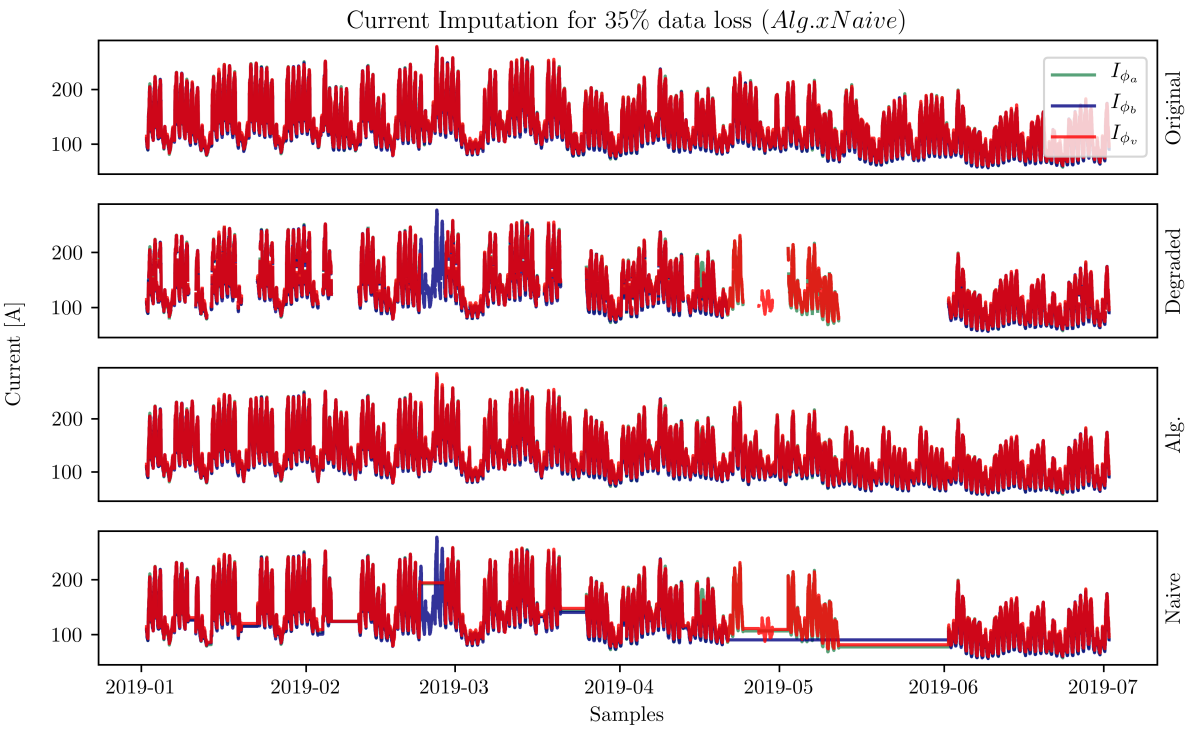

Figure 32: Example of 35% data loss on the current time series.

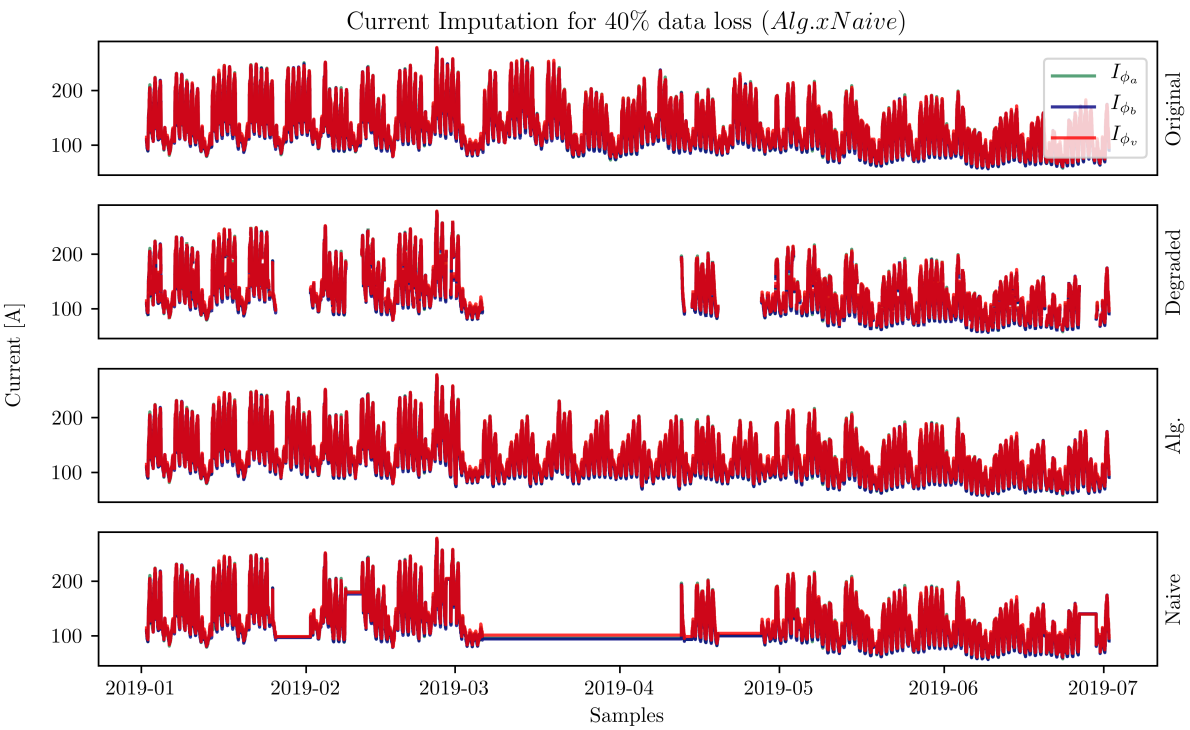

Figure 33: Example of 40% data loss on the current time series.

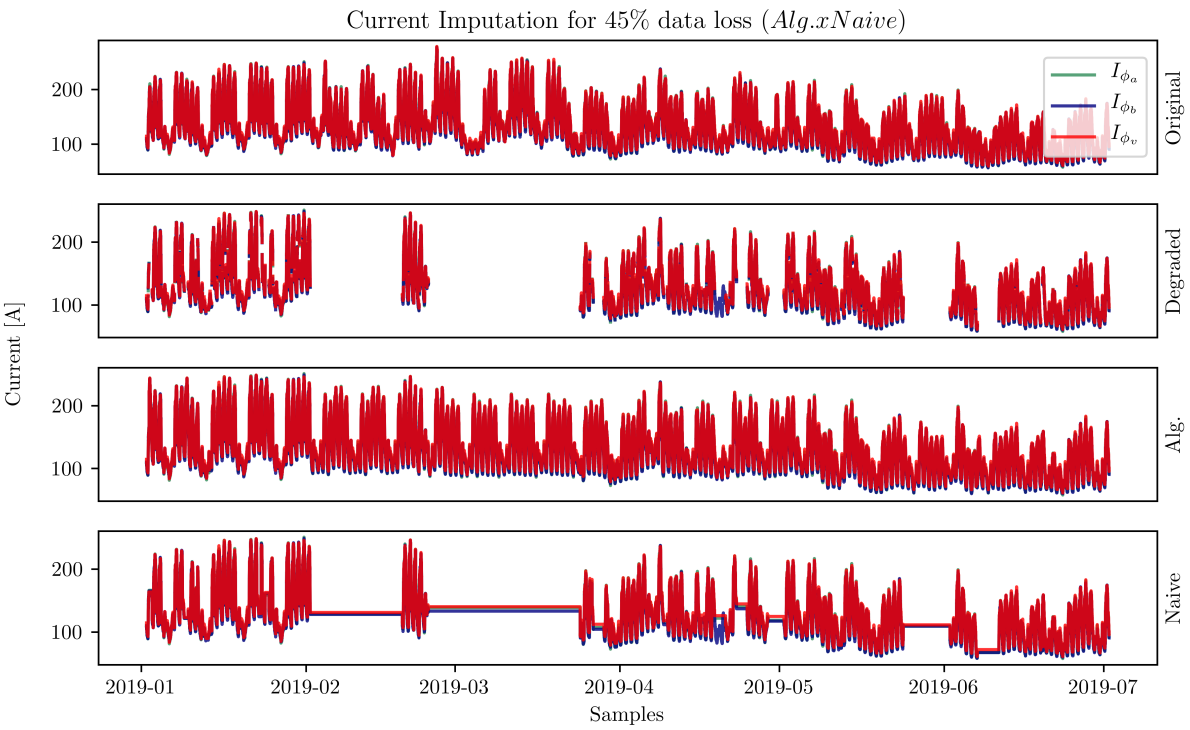

Figure 34: Example of 45% data loss on the current time series.

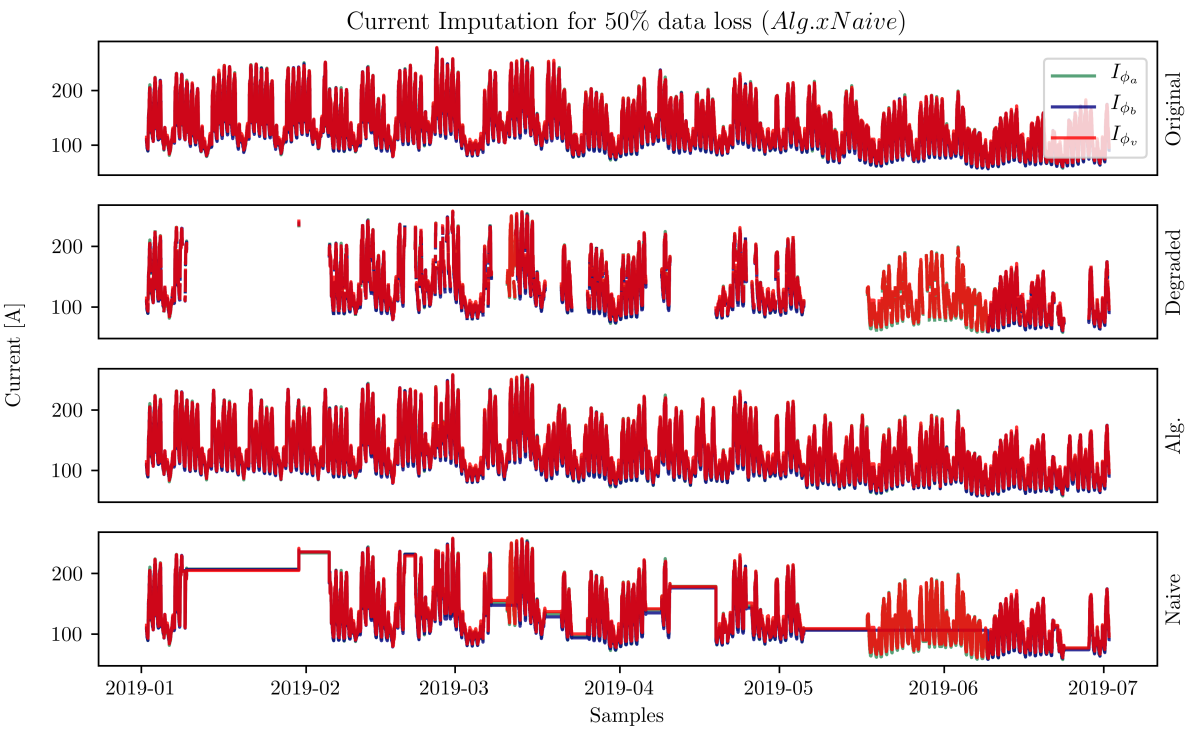

Figure 35: Example of 50% data loss on the current time series.

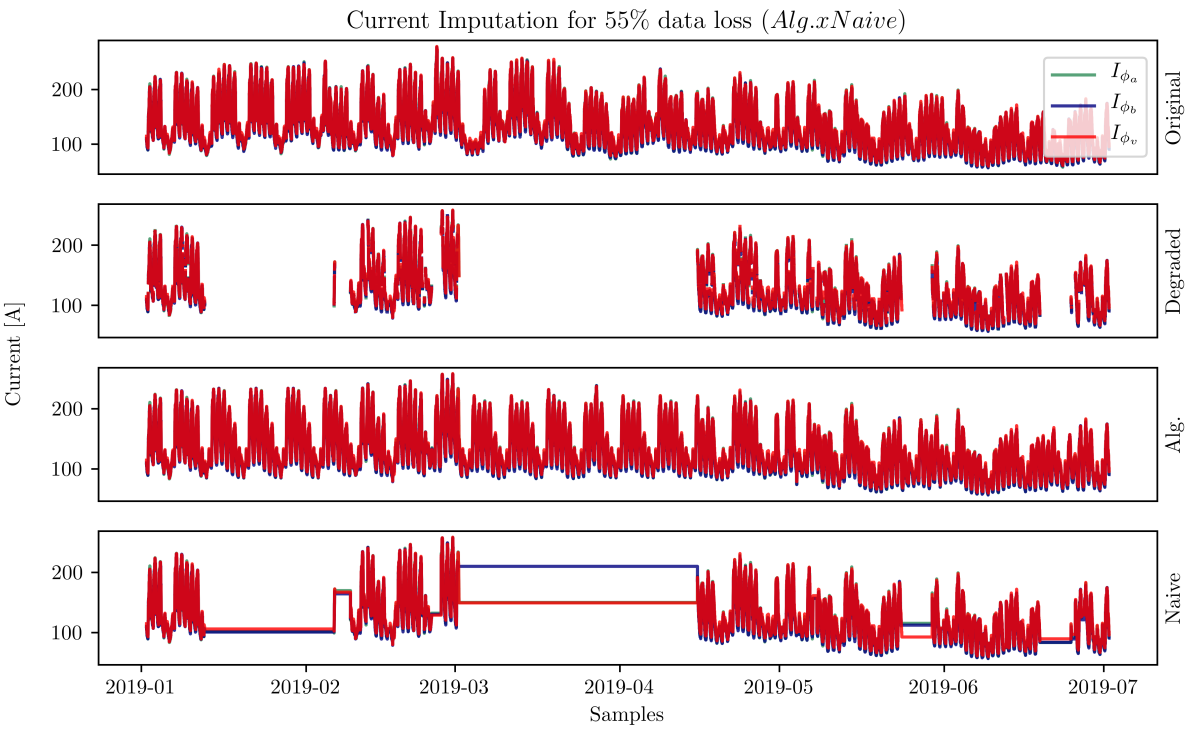

Figure 36: Example of 55% data loss on the current time series.

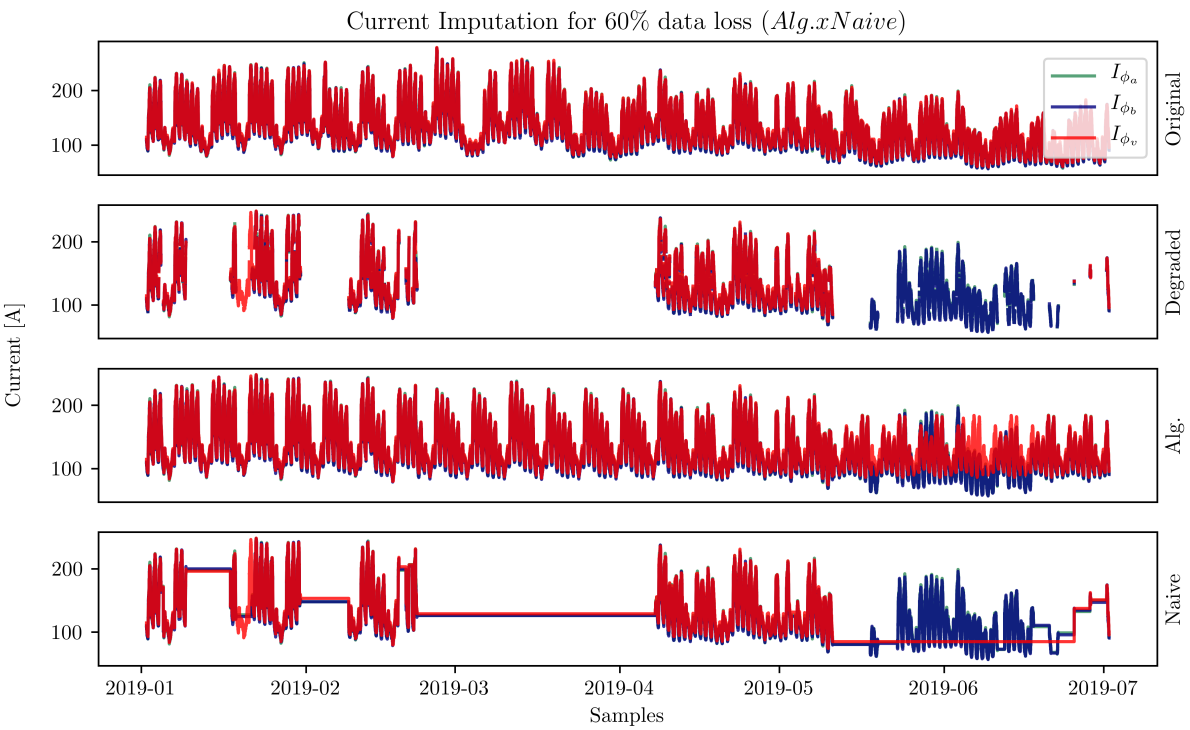

Figure 37: Example of 60% data loss on the current time series.

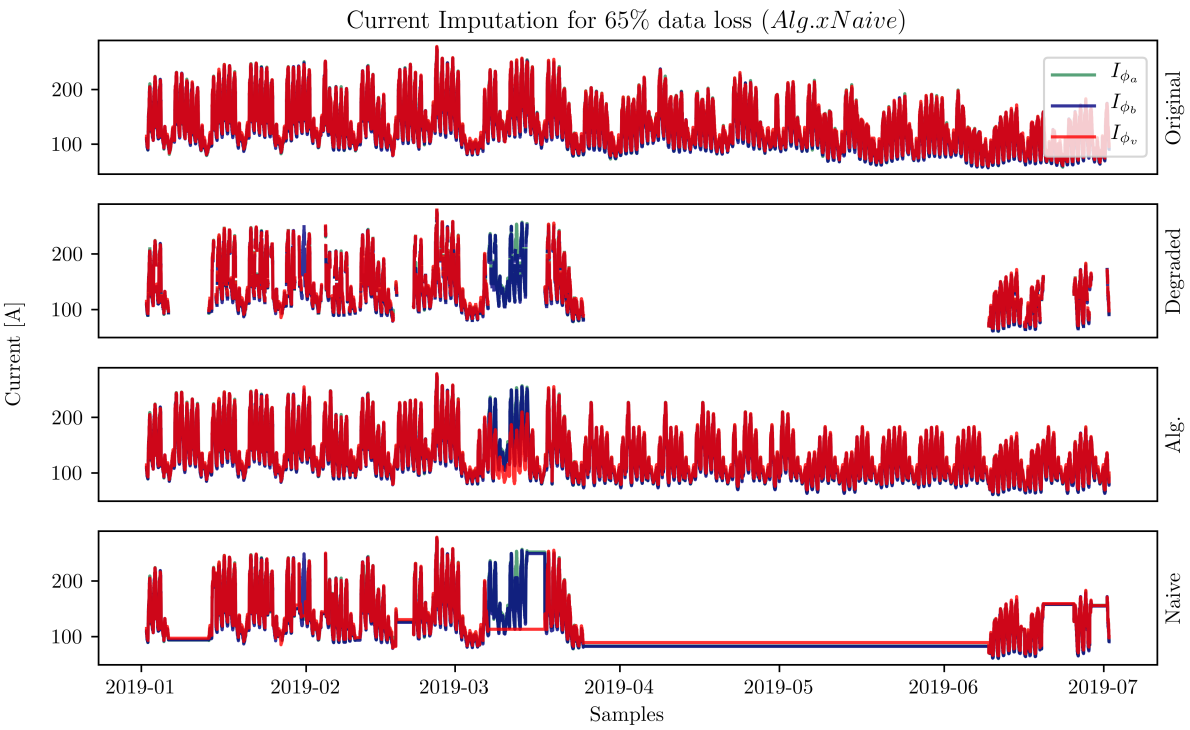

Figure 38: Example of 65% data loss on the current time series.

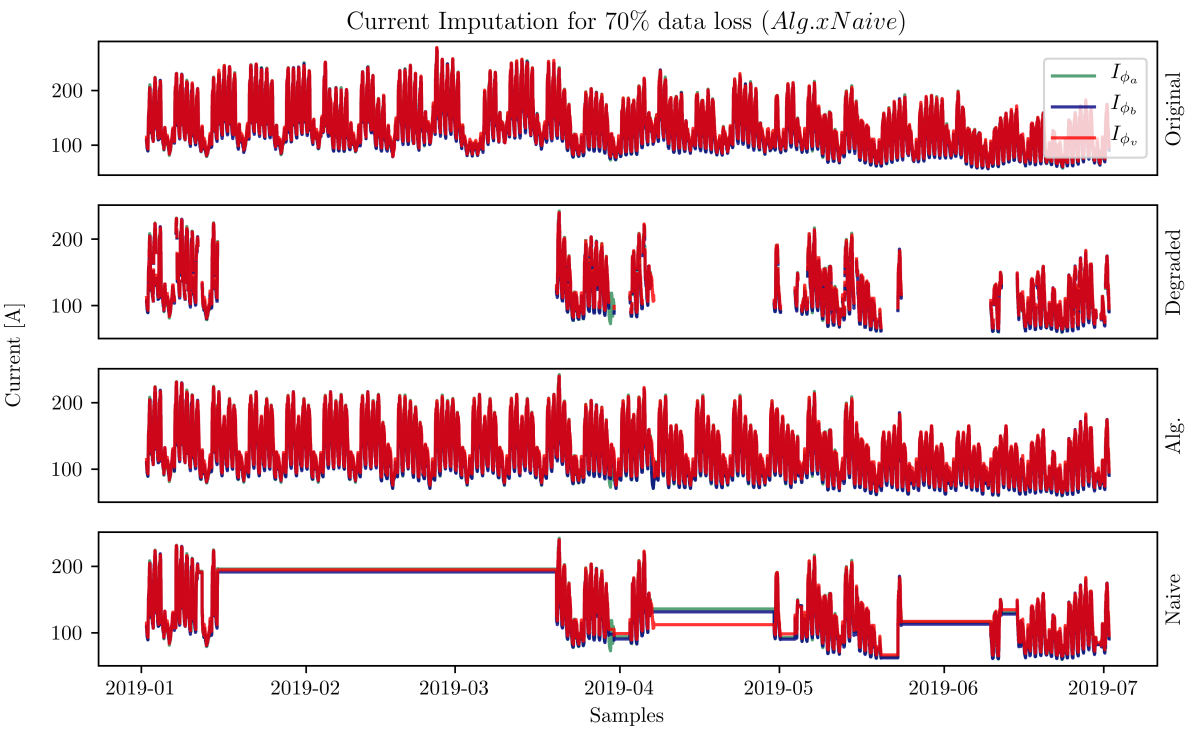

Figure 39: Example of 70% data loss on the current time series.

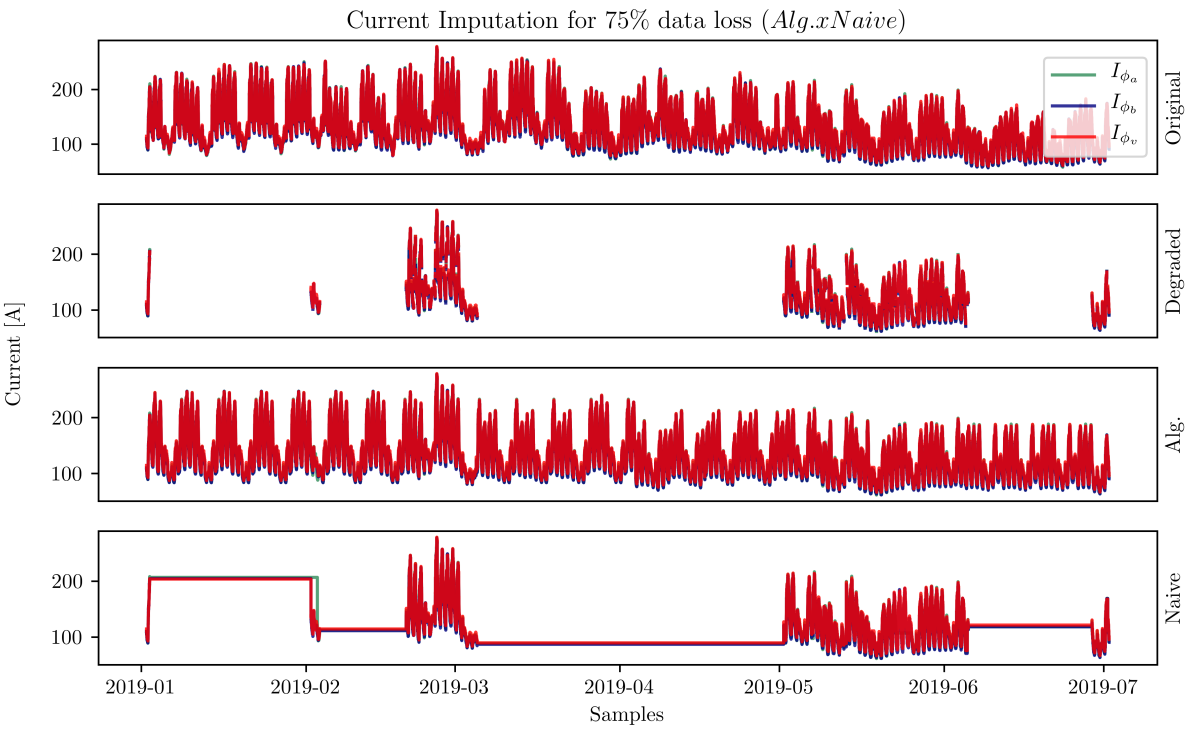

Figure 40: Example of 75% data loss on the current time series.

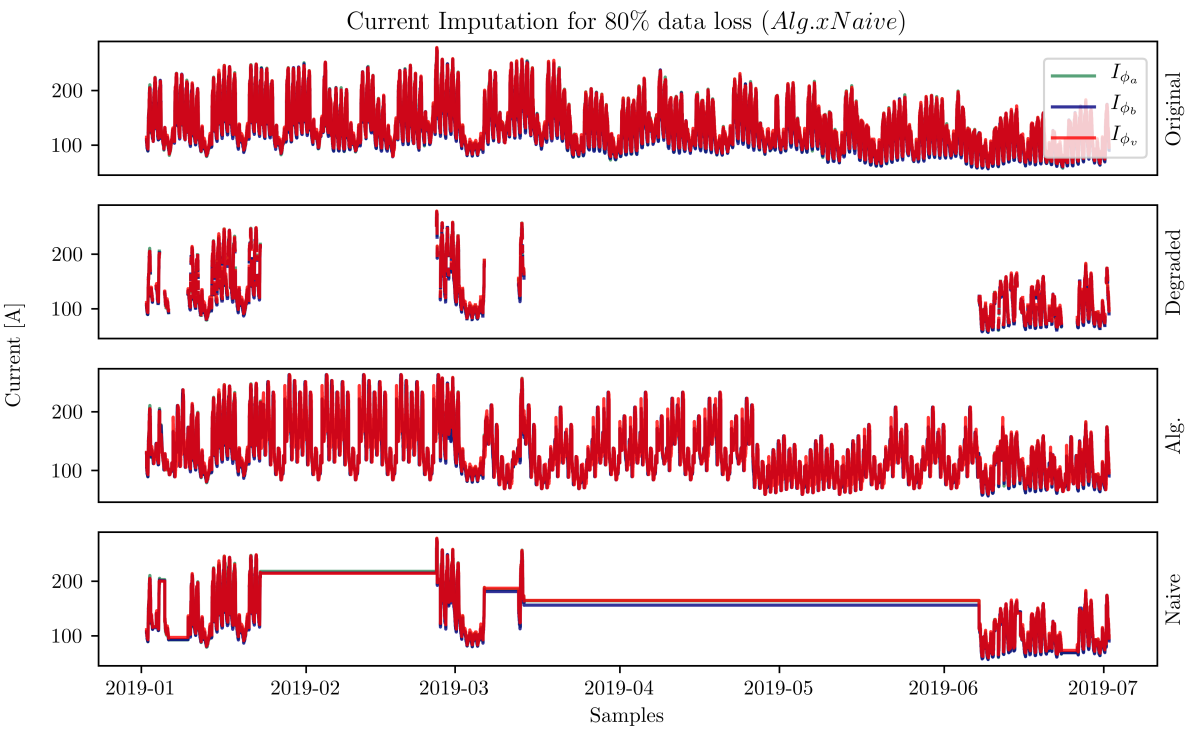

Figure 41: Example of 80% data loss on the current time series.

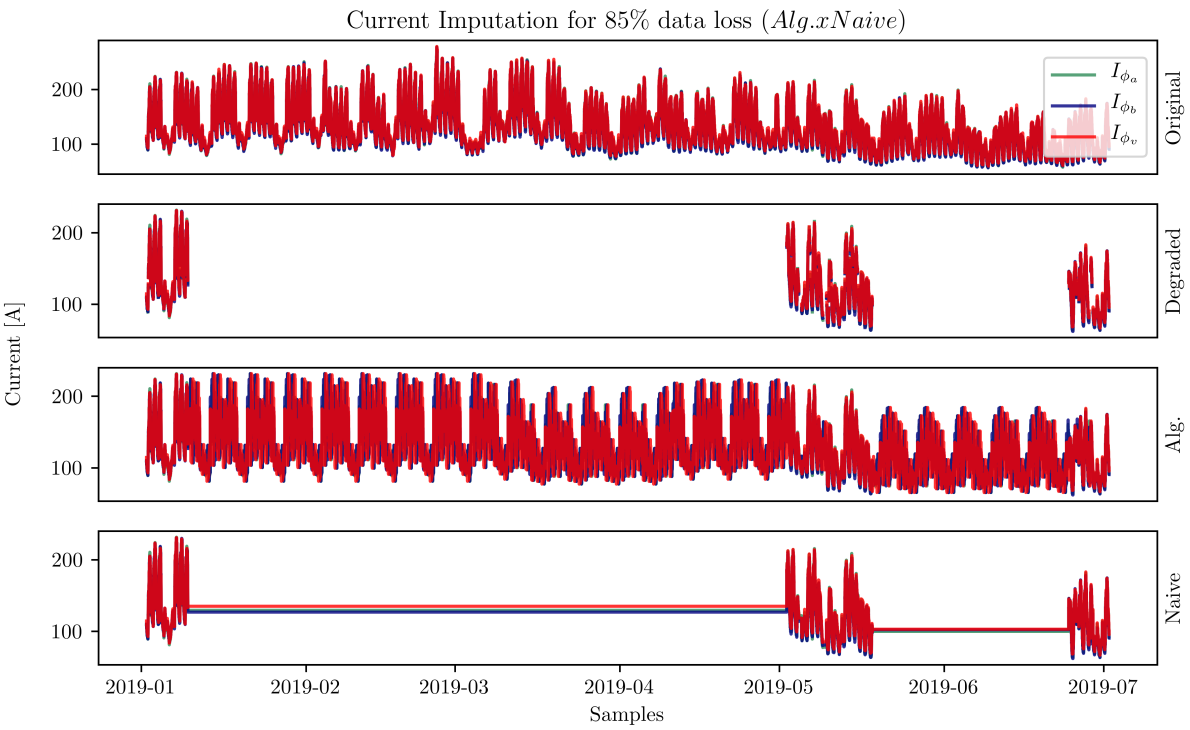

Figure 42: Example of 85% data loss on the current time series.

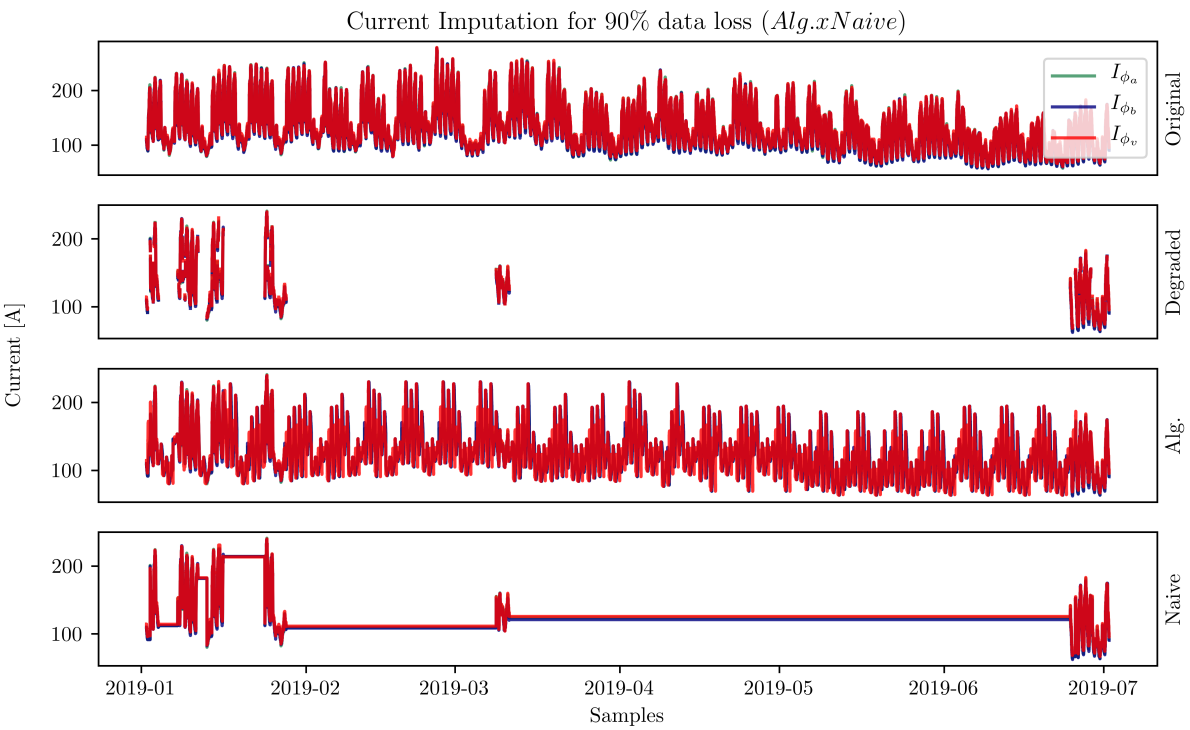

Figure 43: Example of 90% data loss on the current time series.

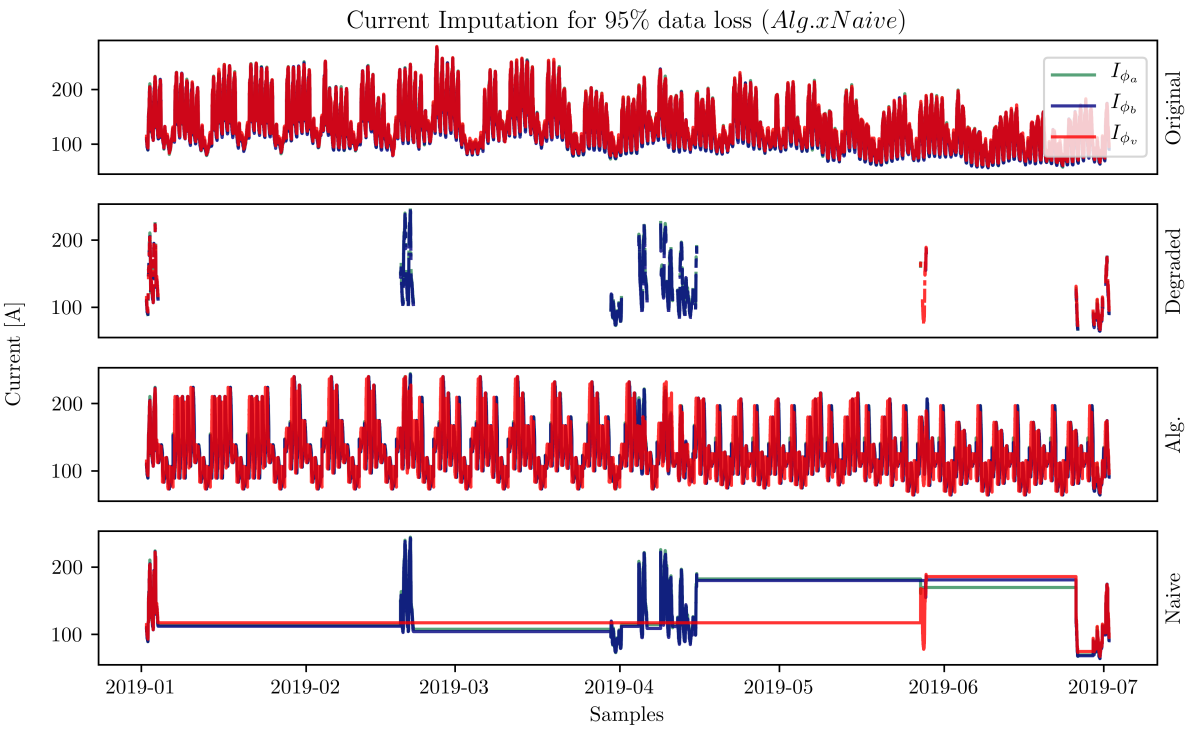

Figure 44: Example of 95% data loss on the current time series.

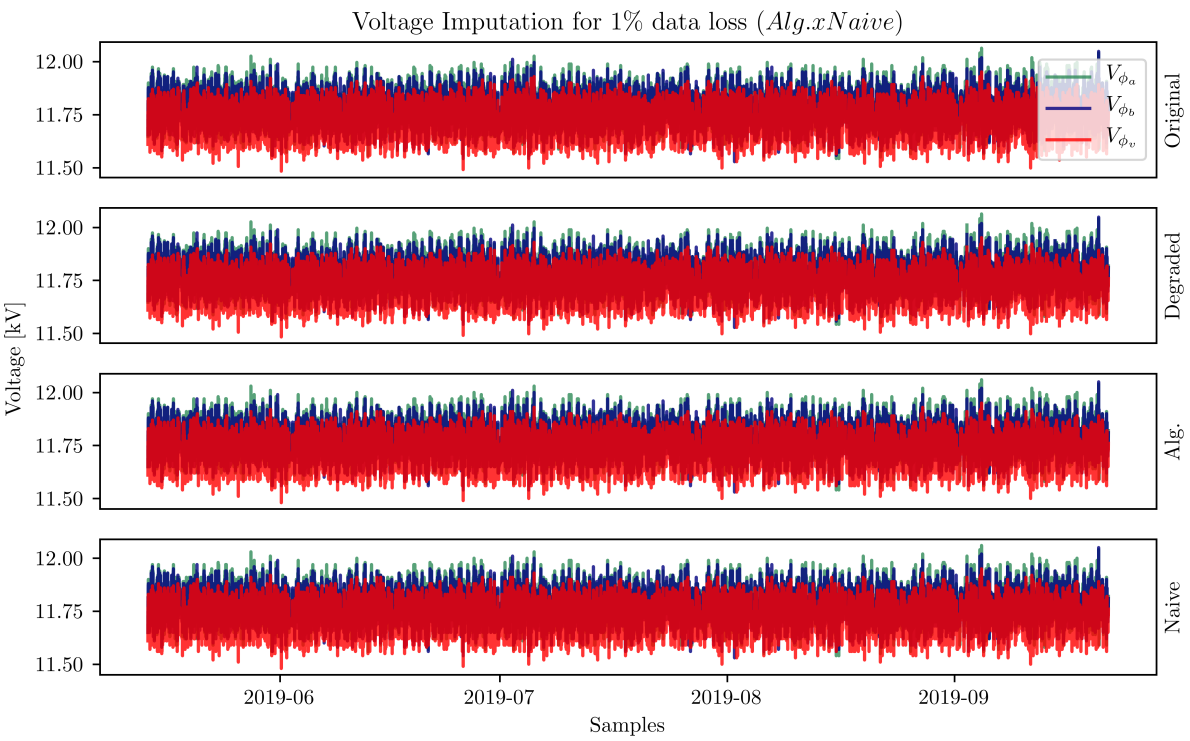

Figure 45: Example of 1% data loss on the voltage time series.

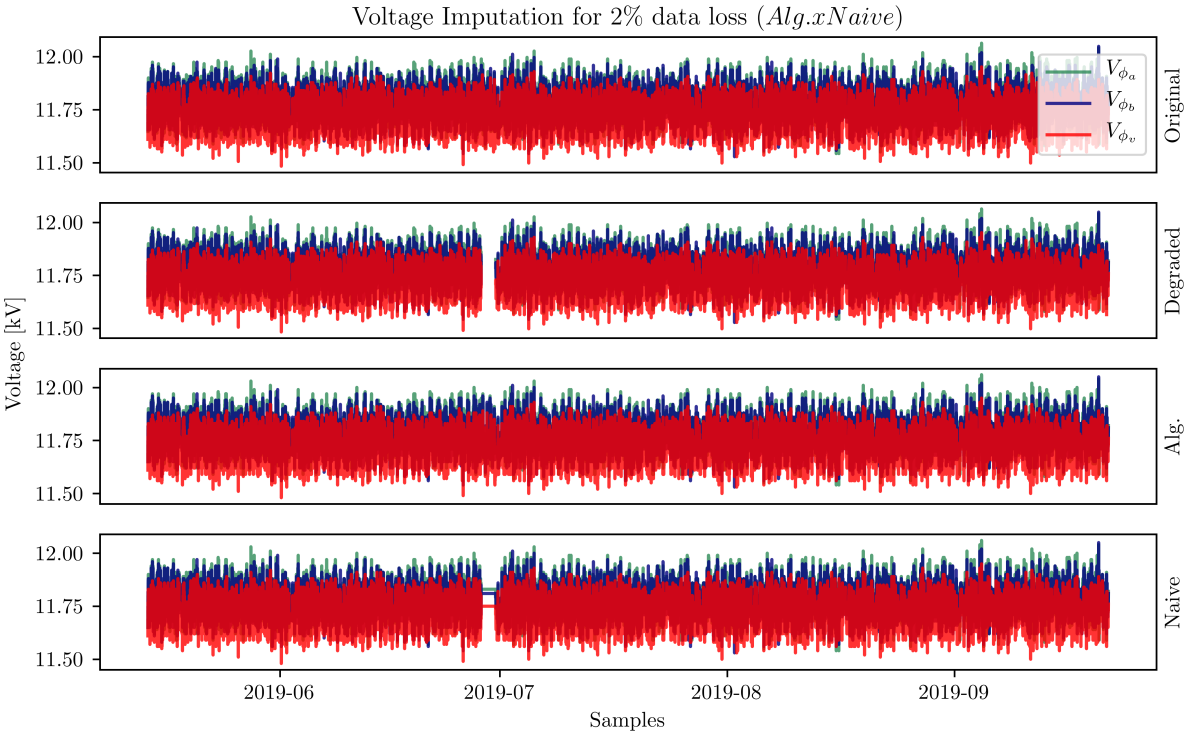

Figure 46: Example of 2% data loss on the voltage time series.

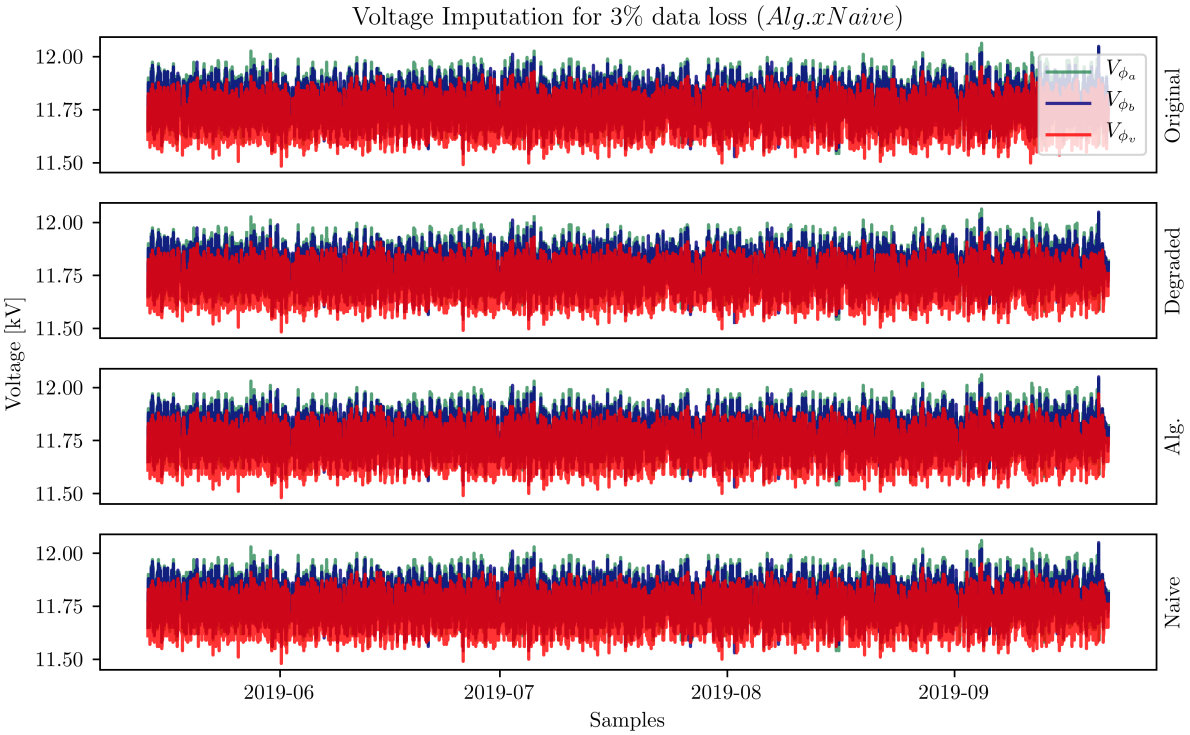

Figure 47: Example of 3% data loss on the voltage time series.

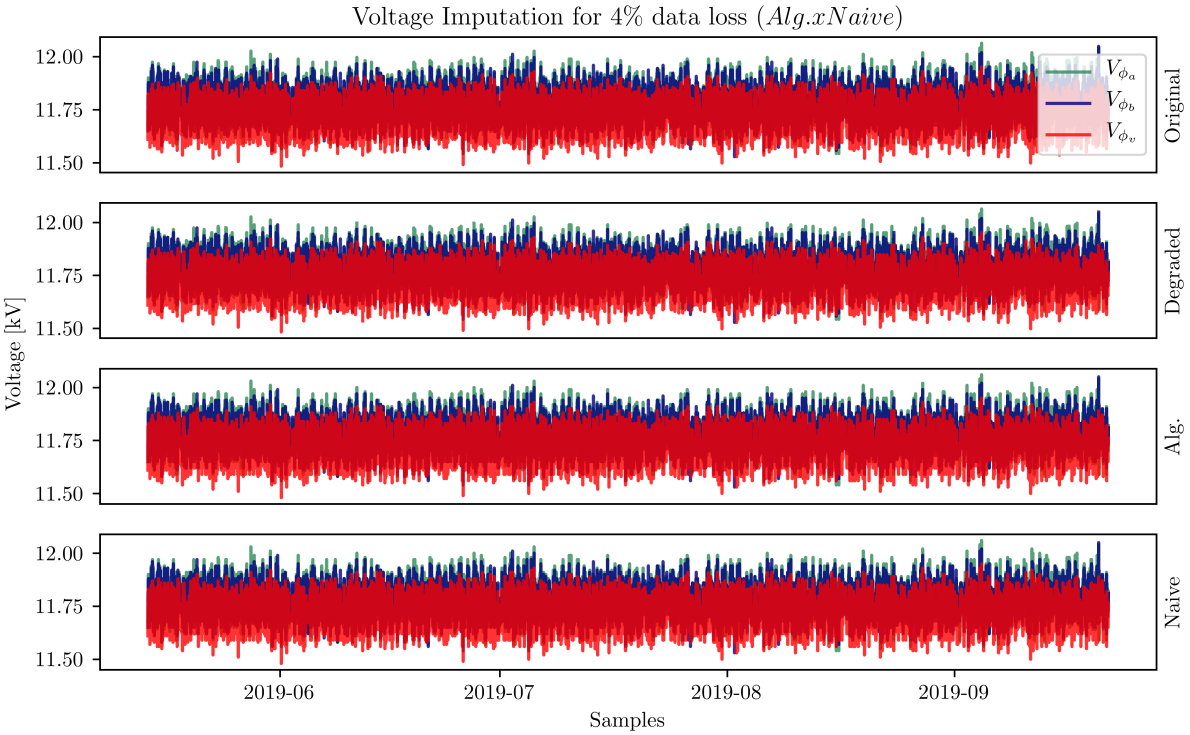

Figure 48: Example of 4% data loss on the voltage time series.

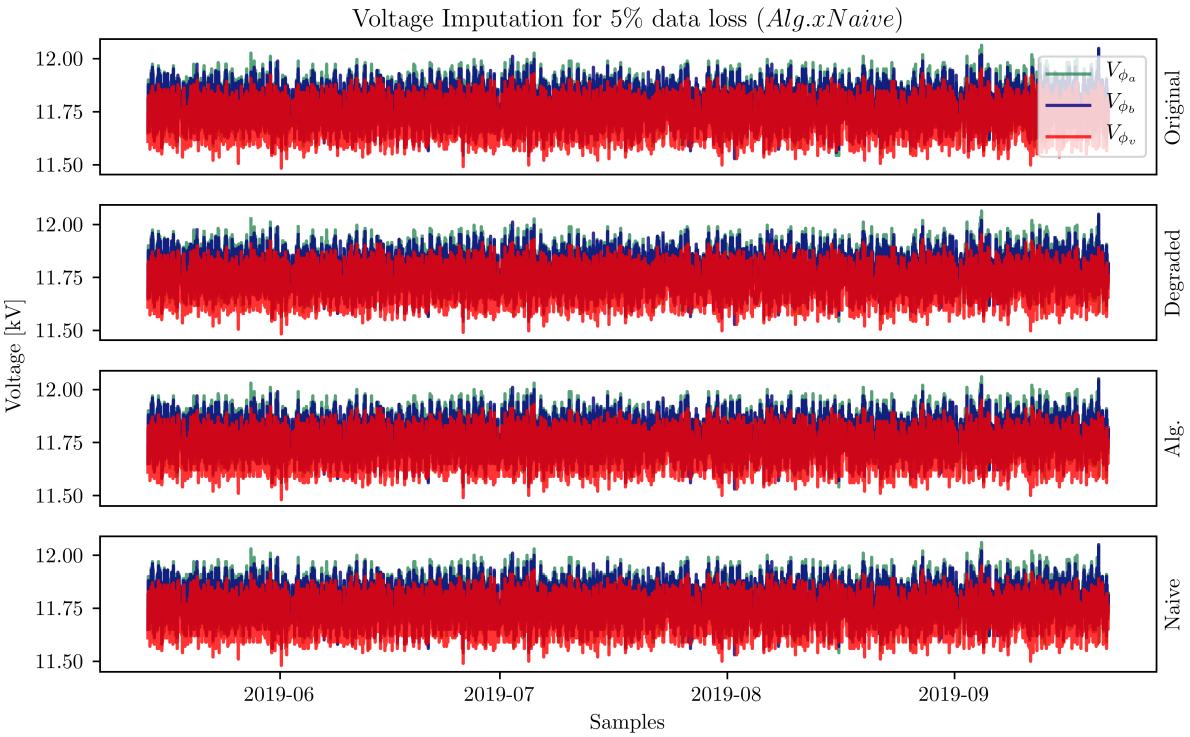

Figure 49: Example of 5% data loss on the voltage time series.

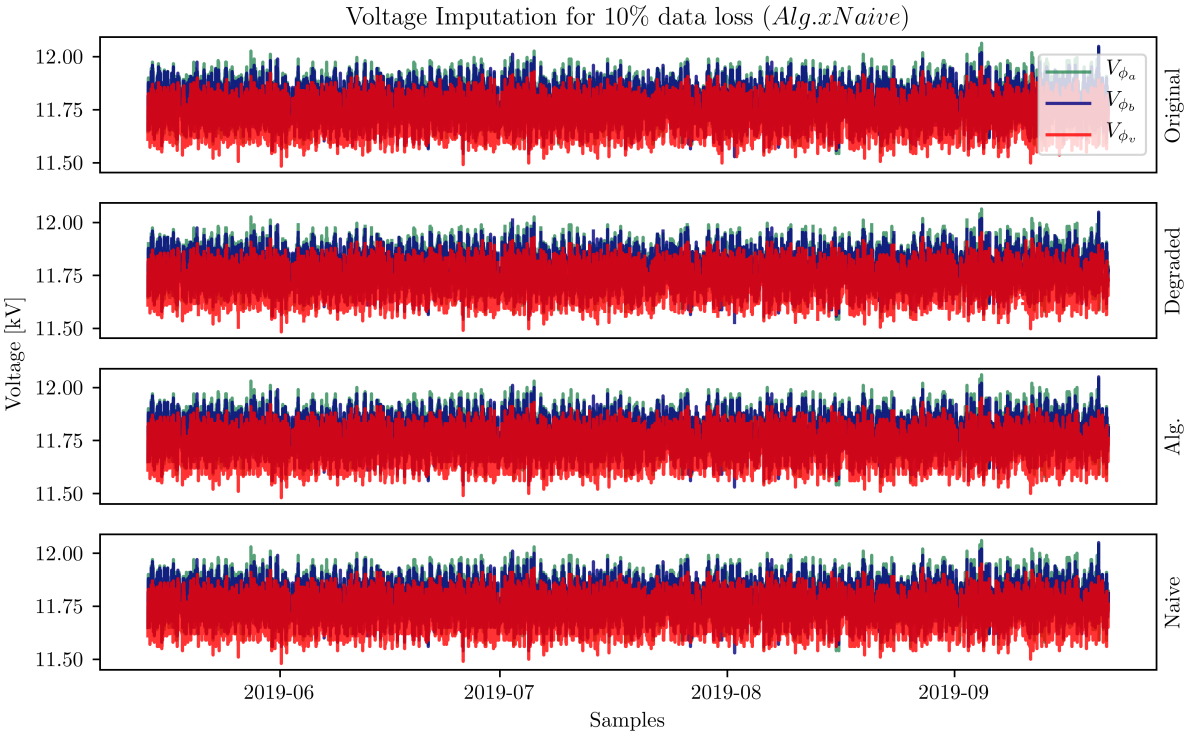

Figure 50: Example of 10% data loss on the voltage time series.

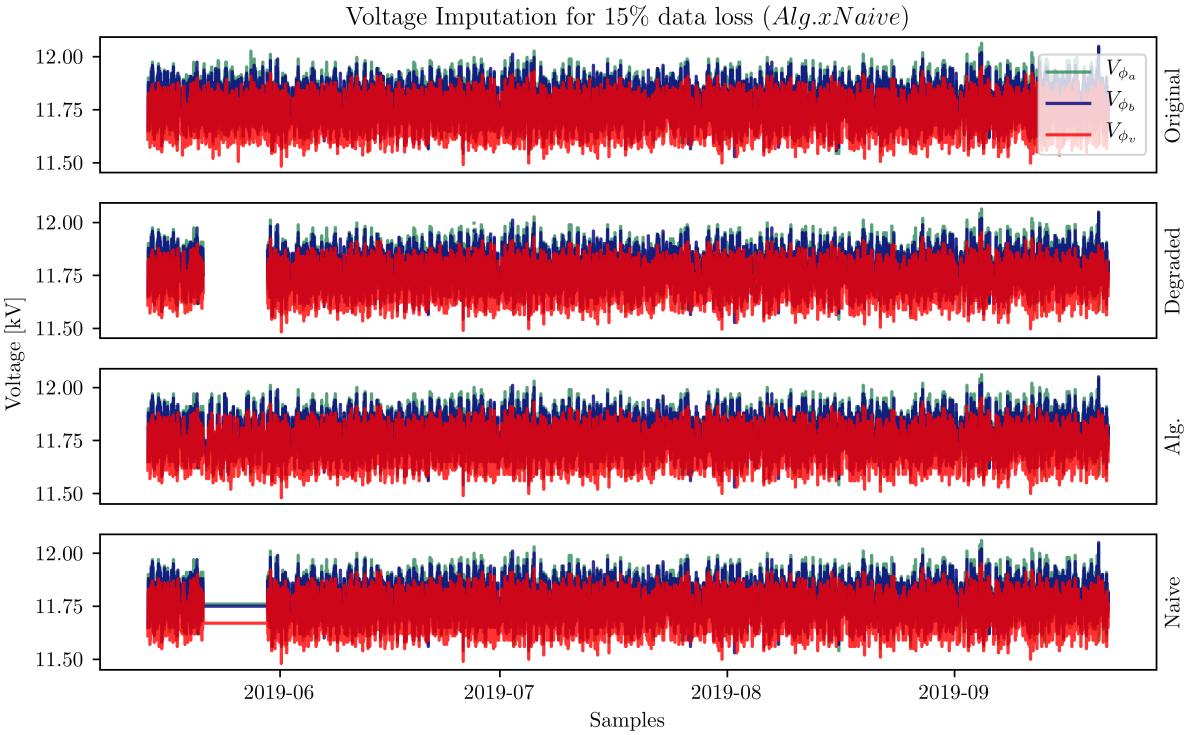

Figure 51: Example of 15% data loss on the voltage time series.

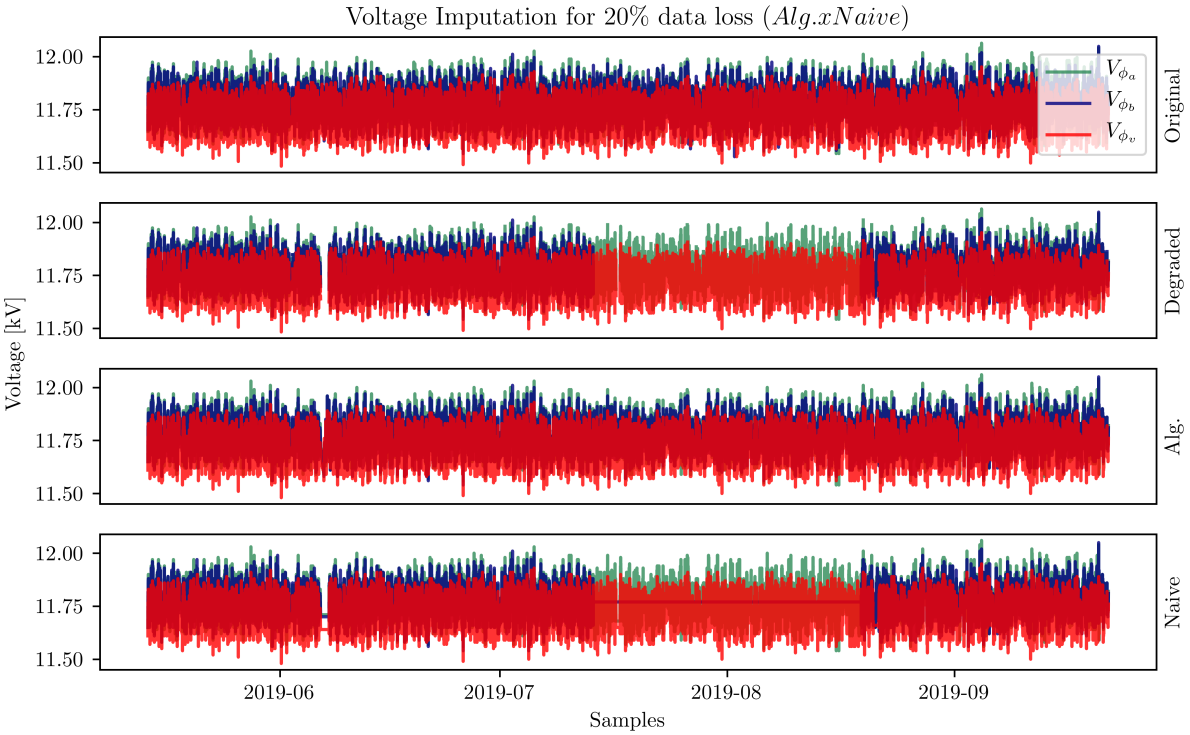

Figure 52: Example of 20% data loss on the voltage time series.

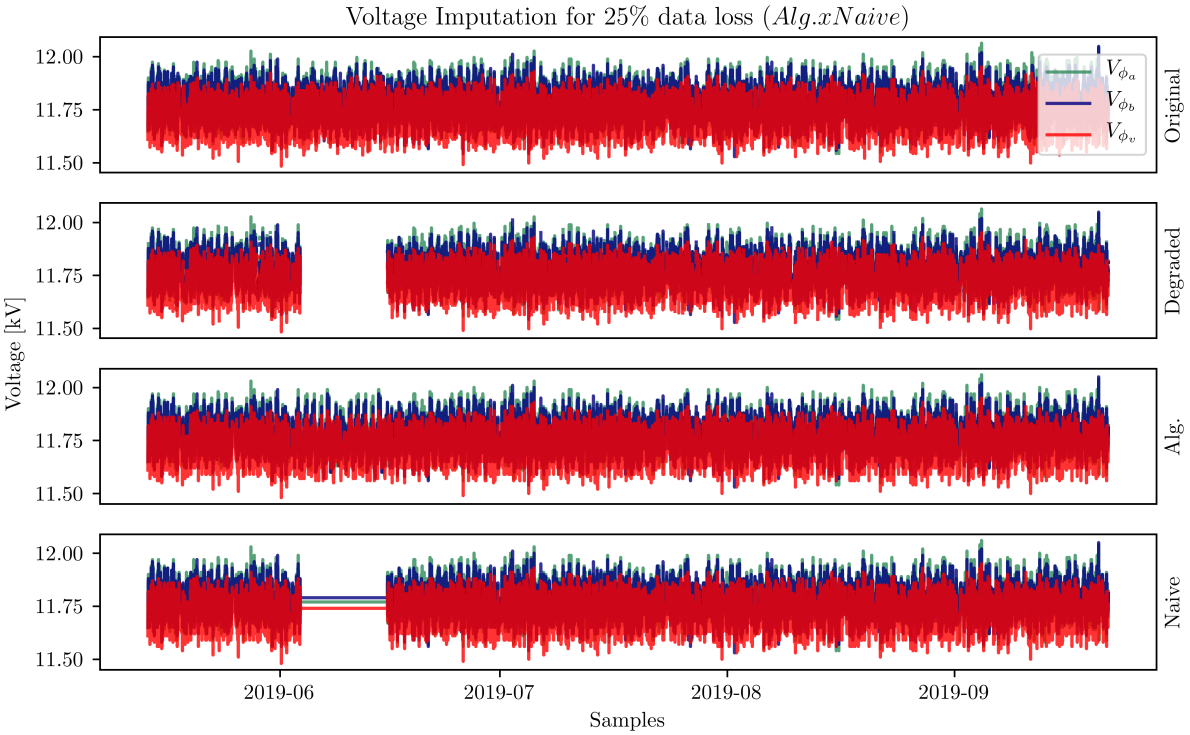

Figure 53: Example of 25% data loss on the voltage time series.

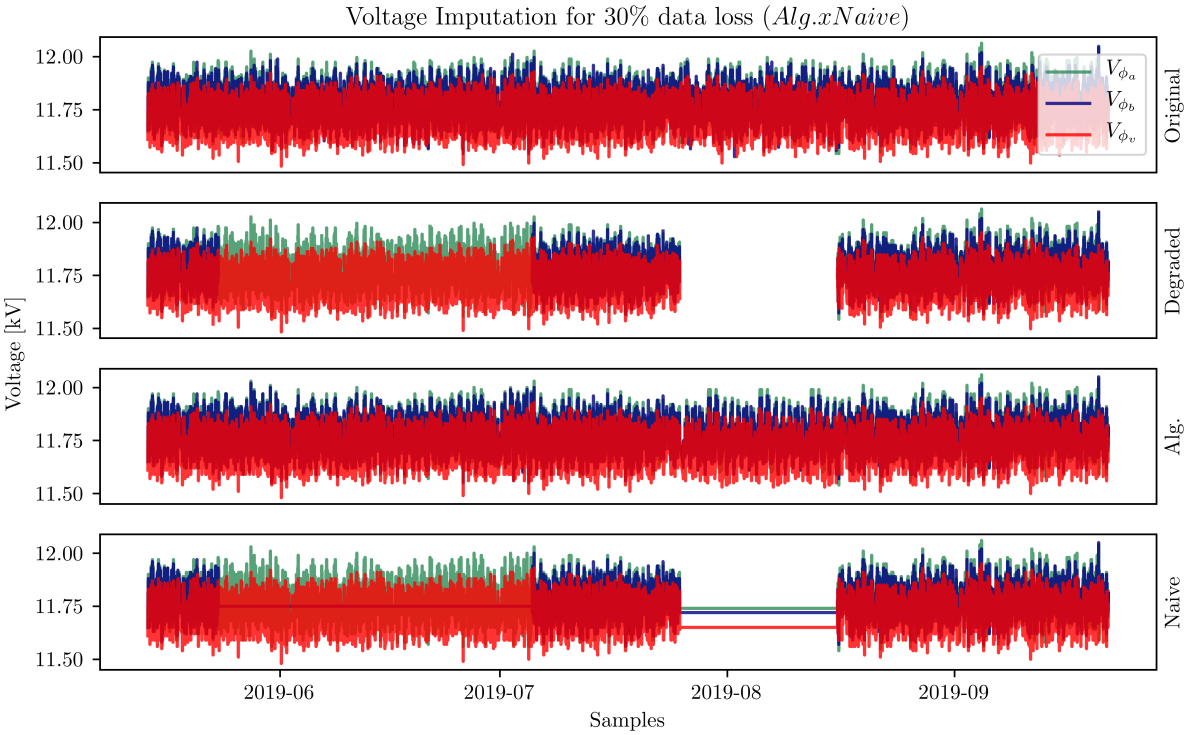

Figure 54: Example of 30% data loss on the voltage time series.

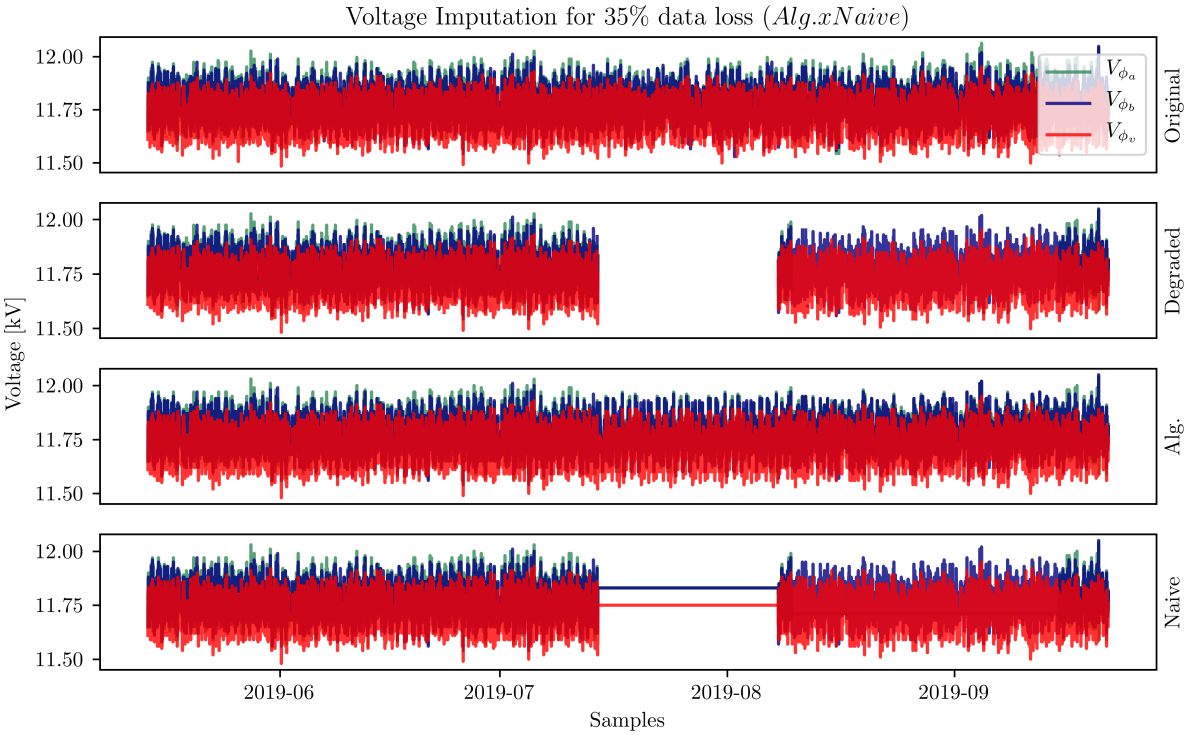

Figure 55: Example of 35% data loss on the voltage time series.

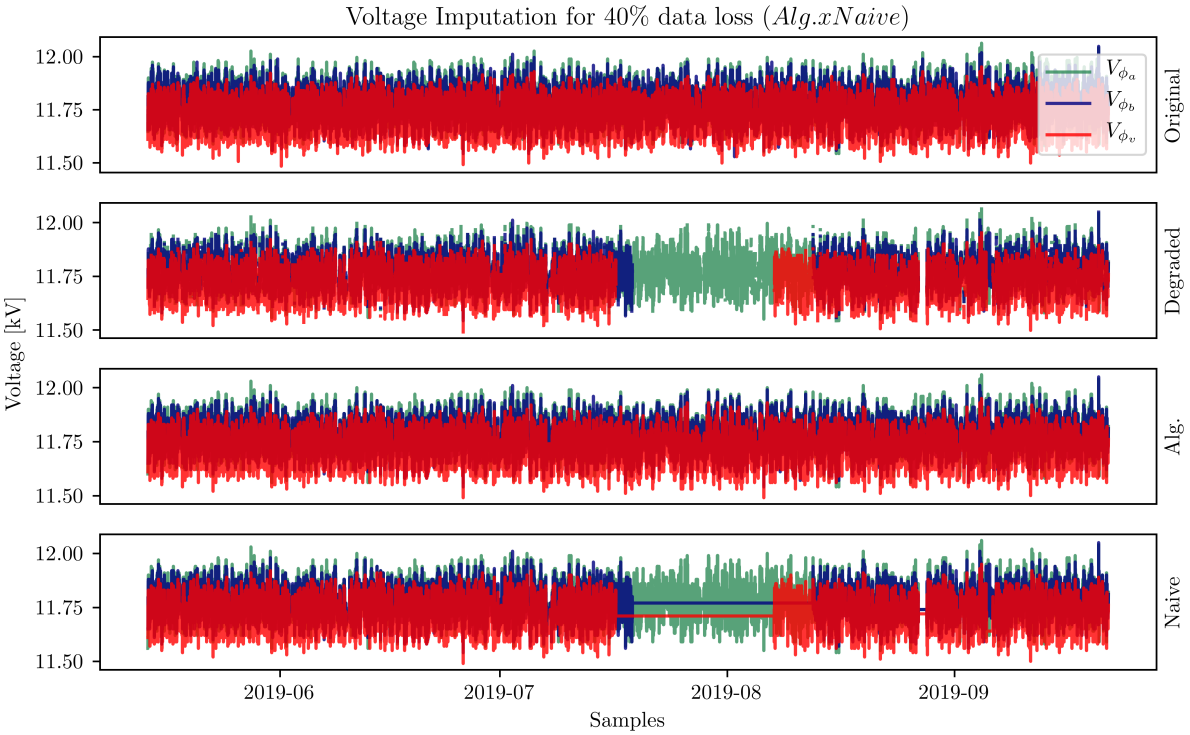

Figure 56: Example of 40% data loss on the voltage time series.

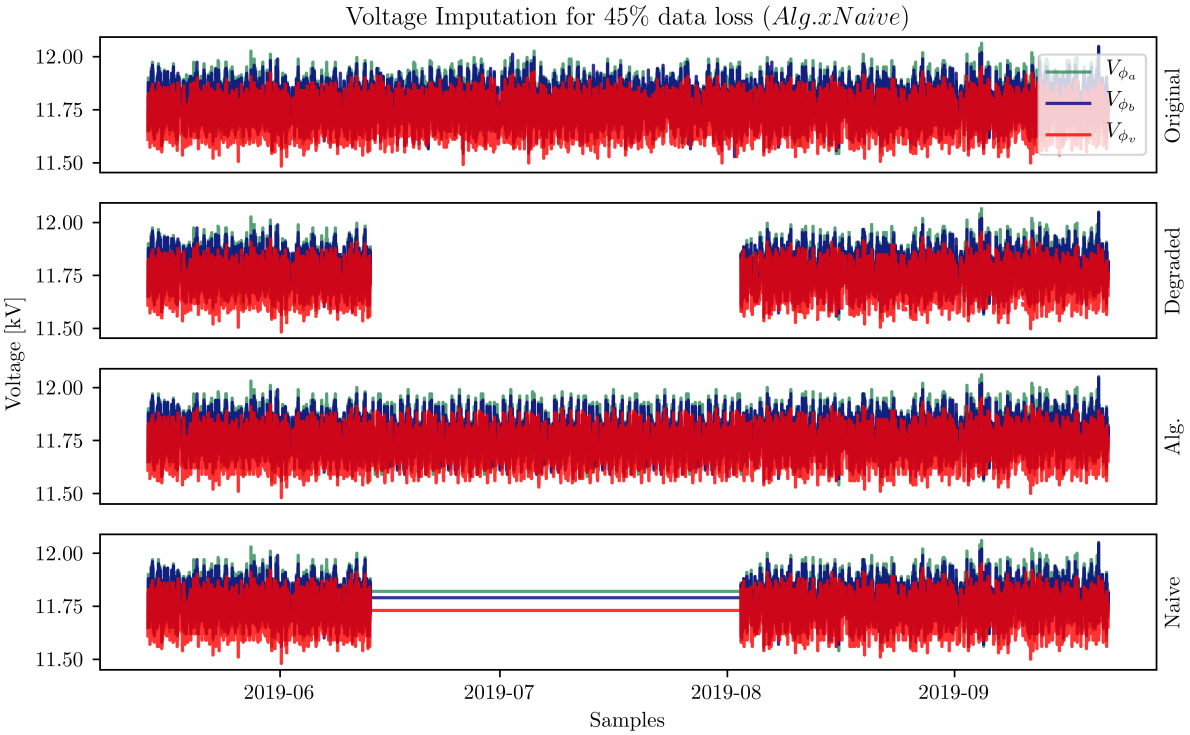

Figure 57: Example of 45% data loss on the voltage time series.

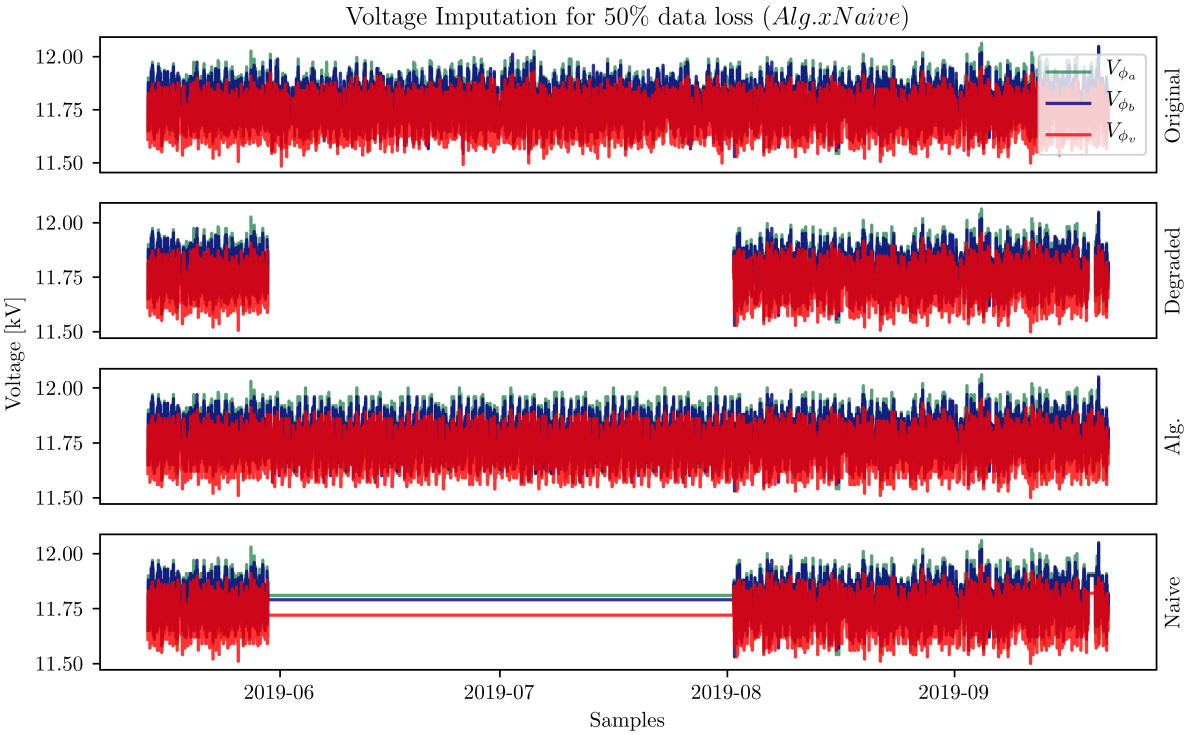

Figure 58: Example of 50% data loss on the voltage time series.

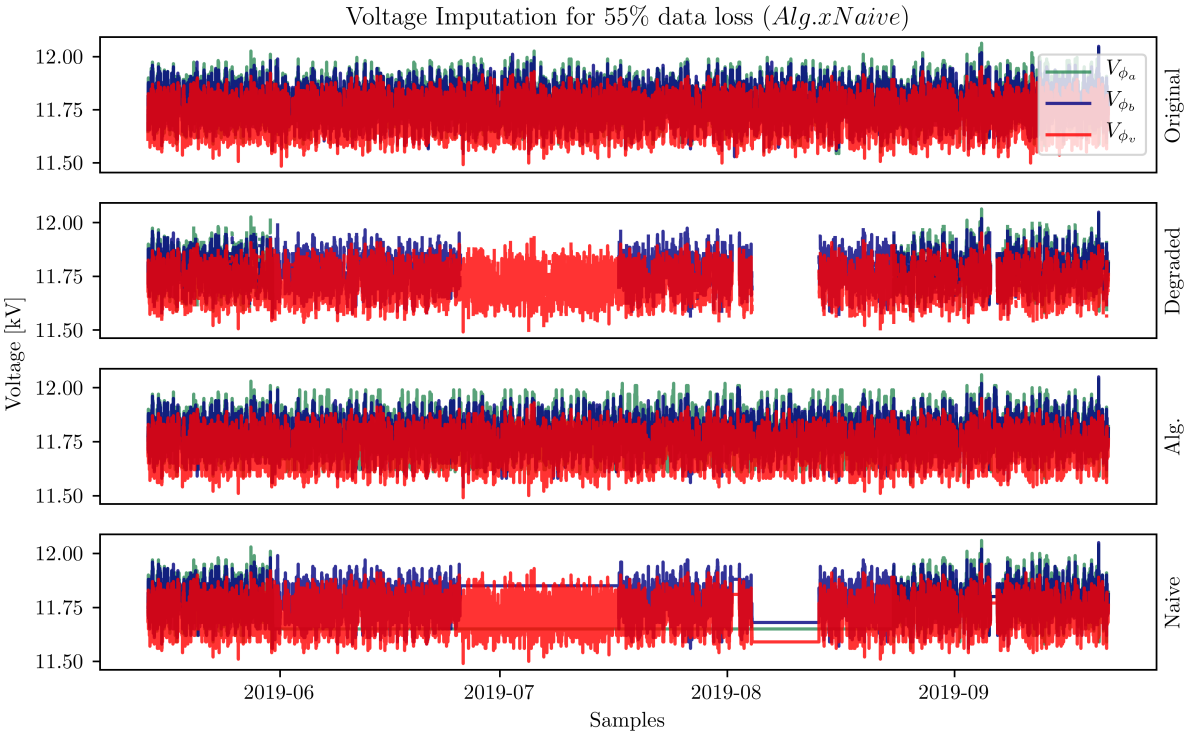

Figure 59: Example of 55% data loss on the voltage time series.

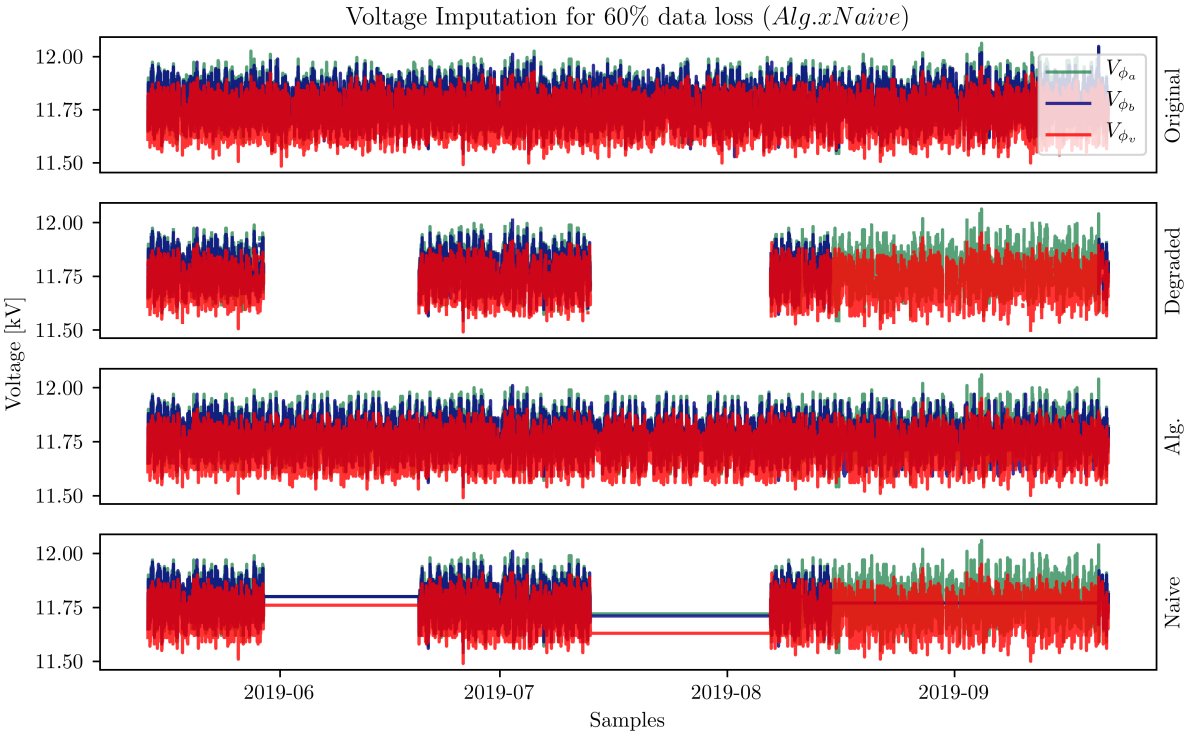

Figure 60: Example of 60% data loss on the voltage time series.

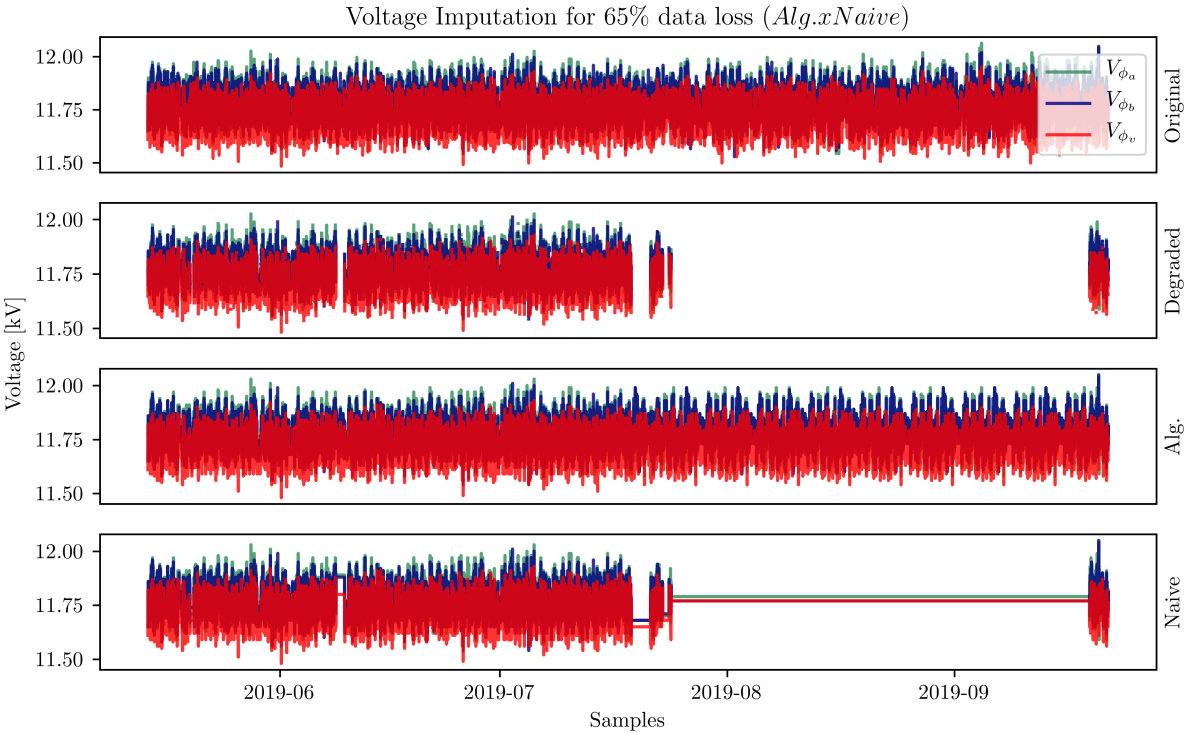

Figure 61: Example of 65% data loss on the voltage time series.

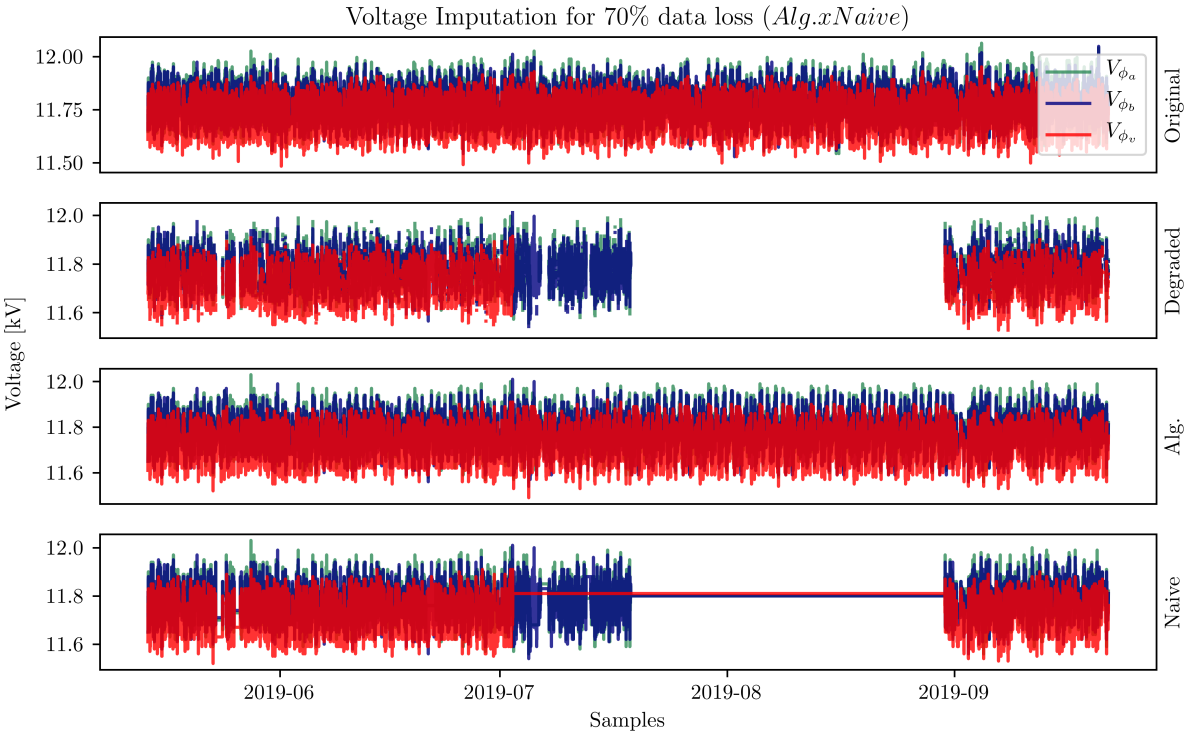

Figure 62: Example of 70% data loss on the voltage time series.

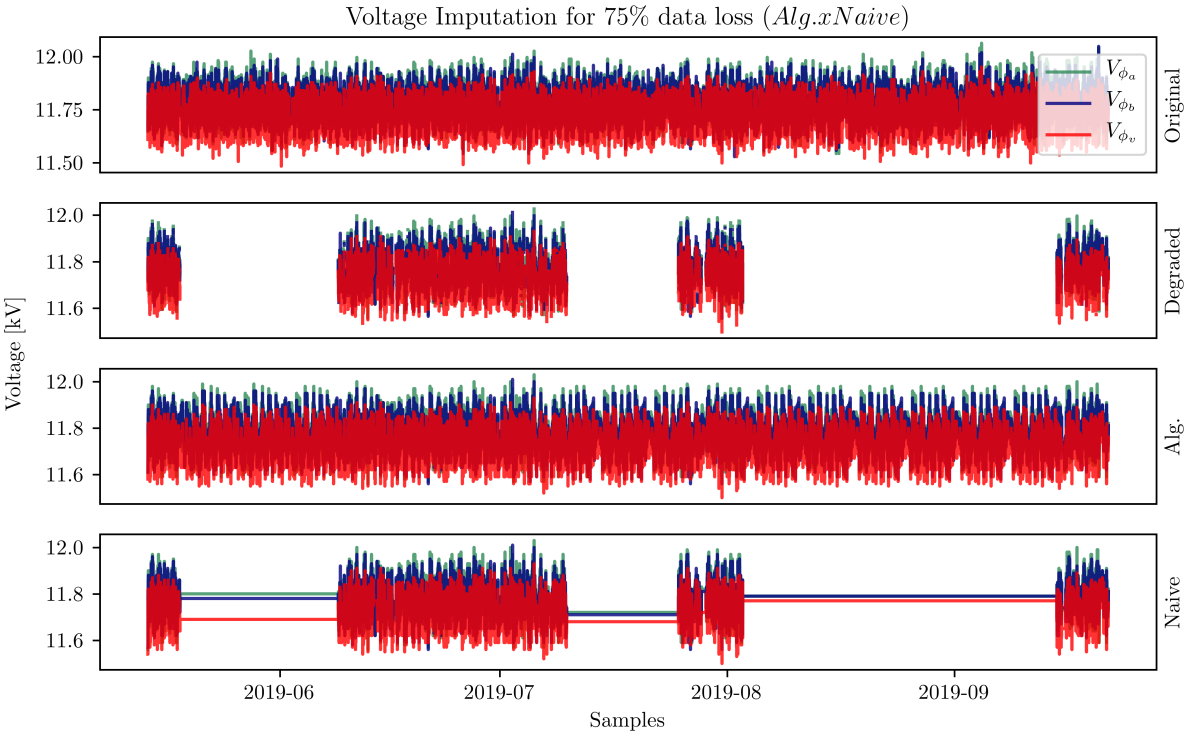

Figure 63: Example of 75% data loss on the voltage time series.

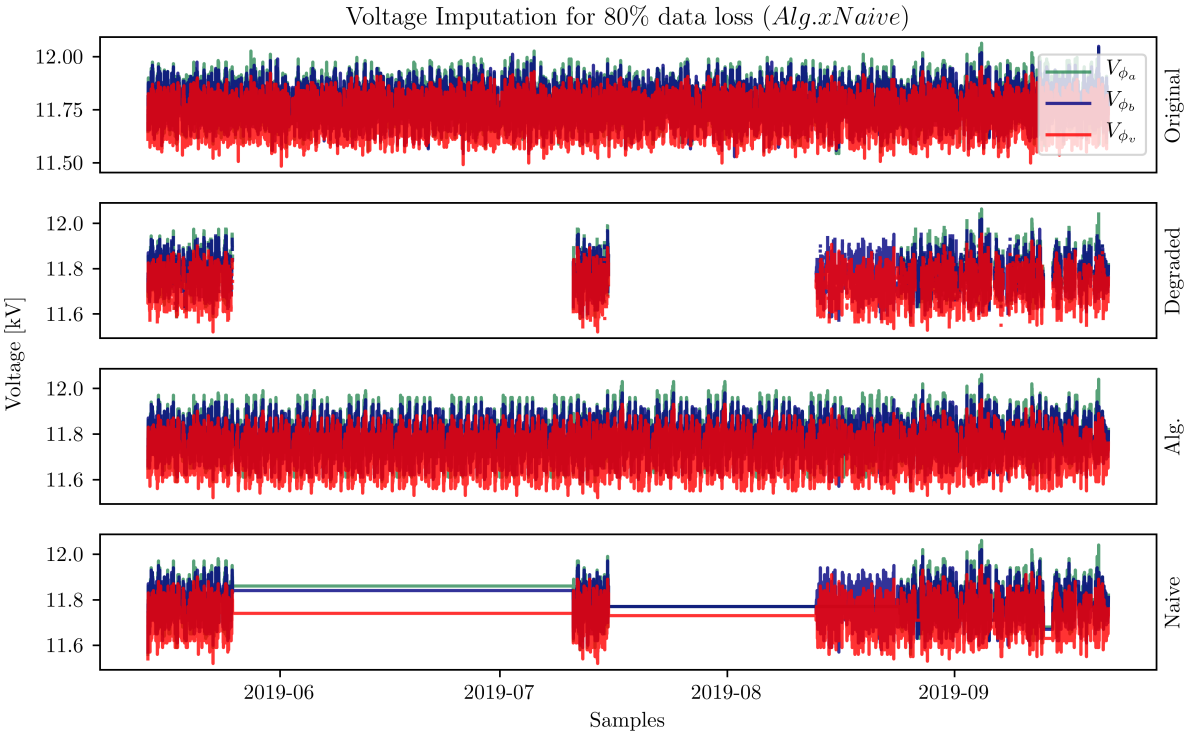

Figure 64: Example of 80% data loss on the voltage time series.

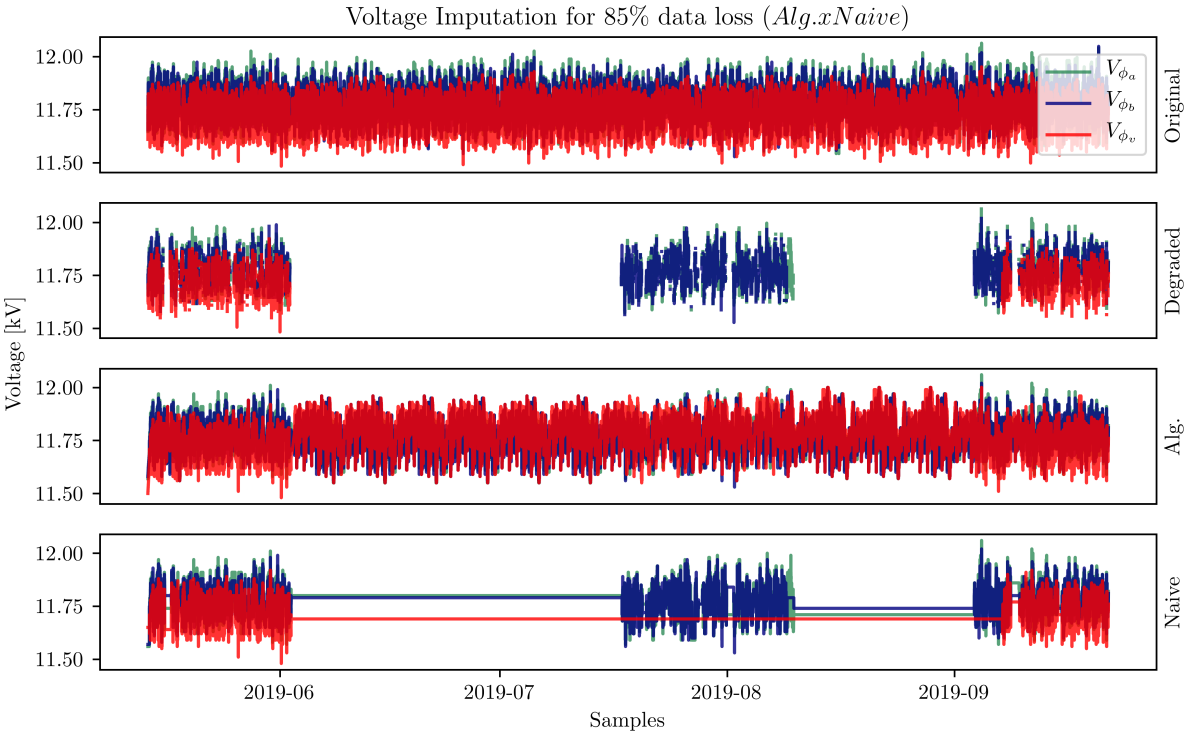

Figure 65: Example of 85% data loss on the voltage time series.

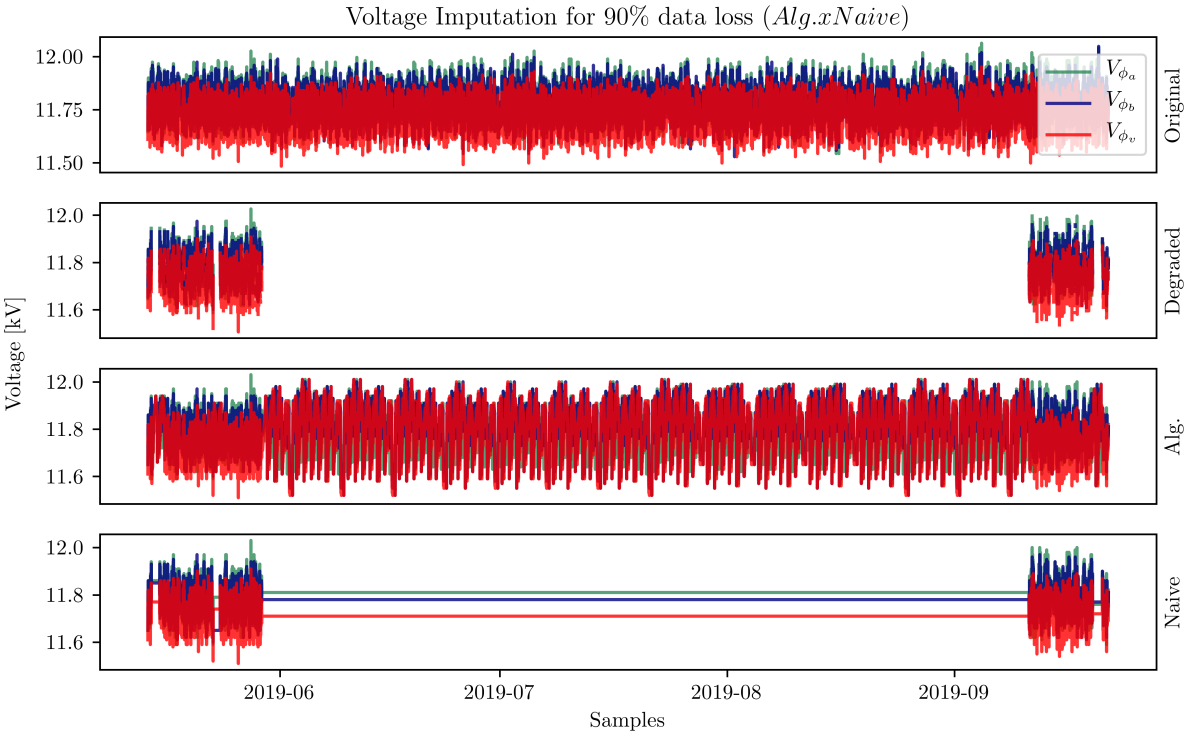

Figure 66: Example of 90% data loss on the voltage time series.

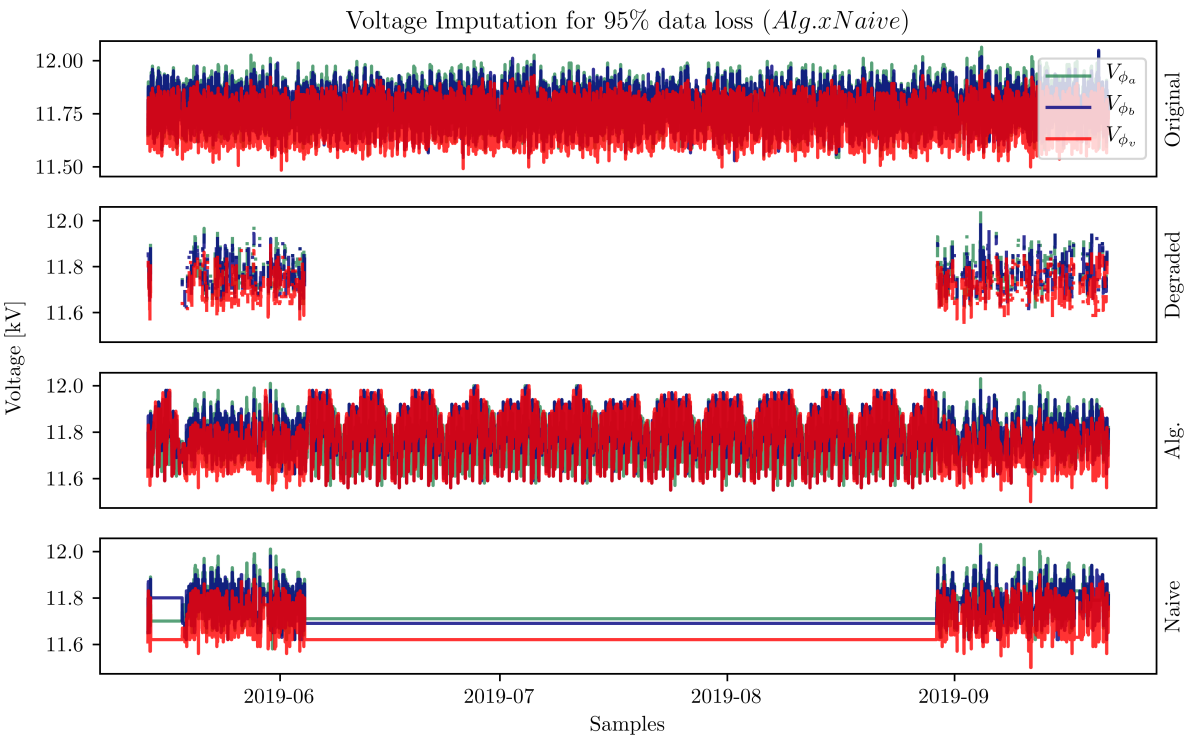

Figure 67: Example of 95% data loss on the voltage time series.

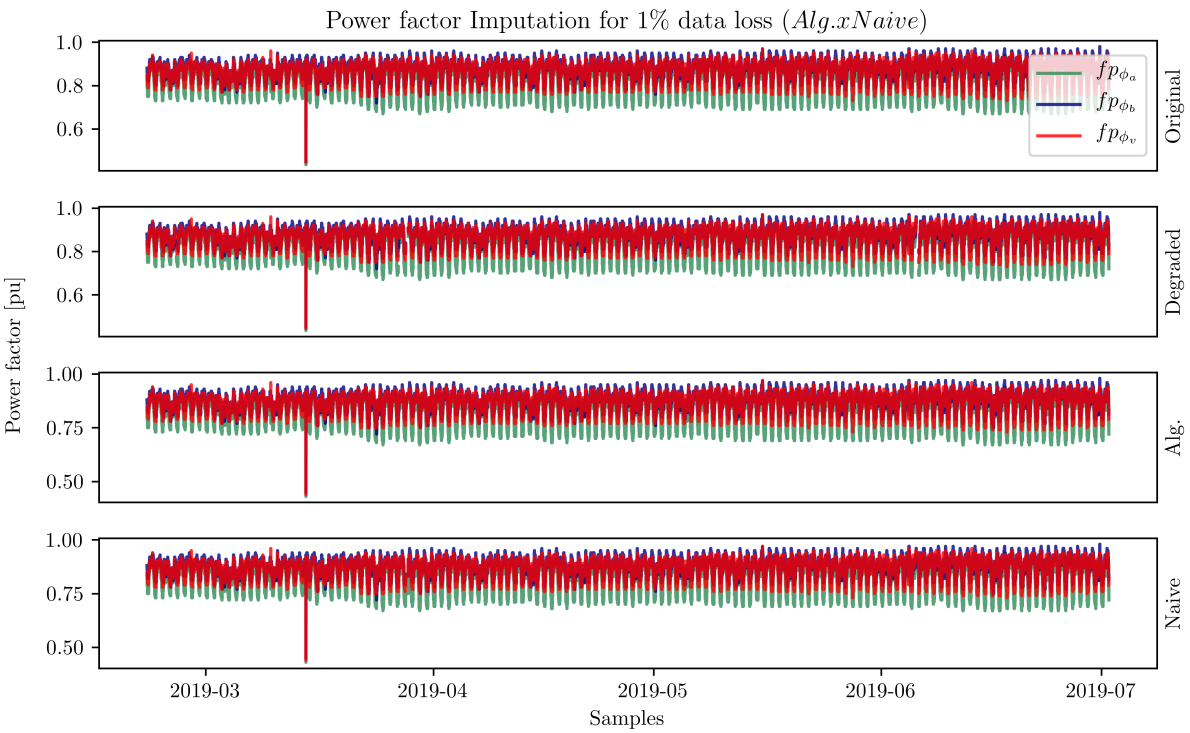

Figure 68: Example of 1% data loss on the power factor time series.

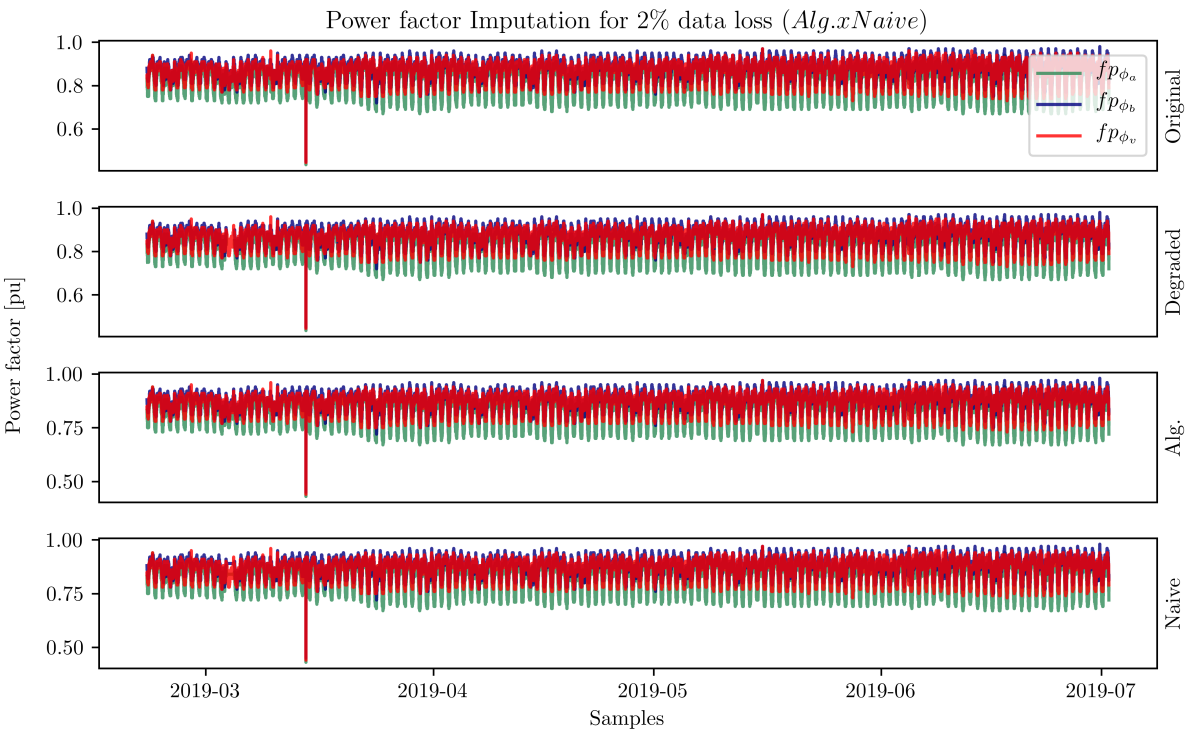

Figure 69: Example of 2% data loss on the power factor time series.

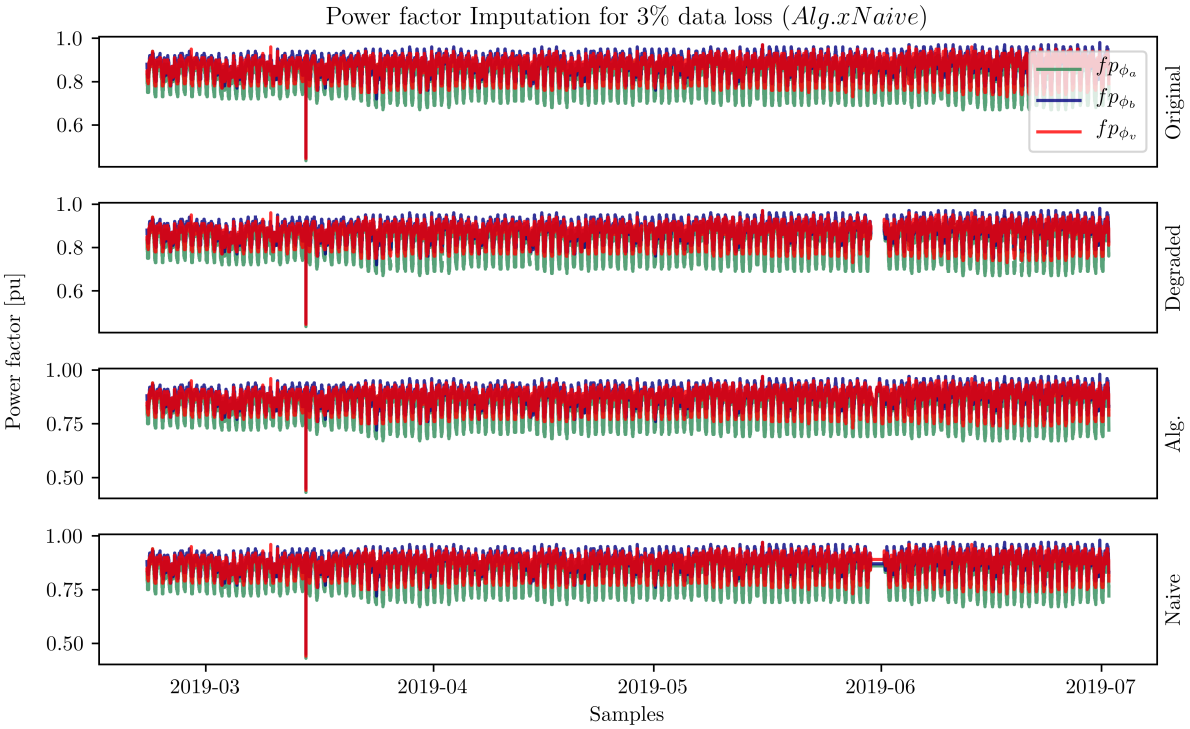

Figure 70: Example of 3% data loss on the power factor time series.

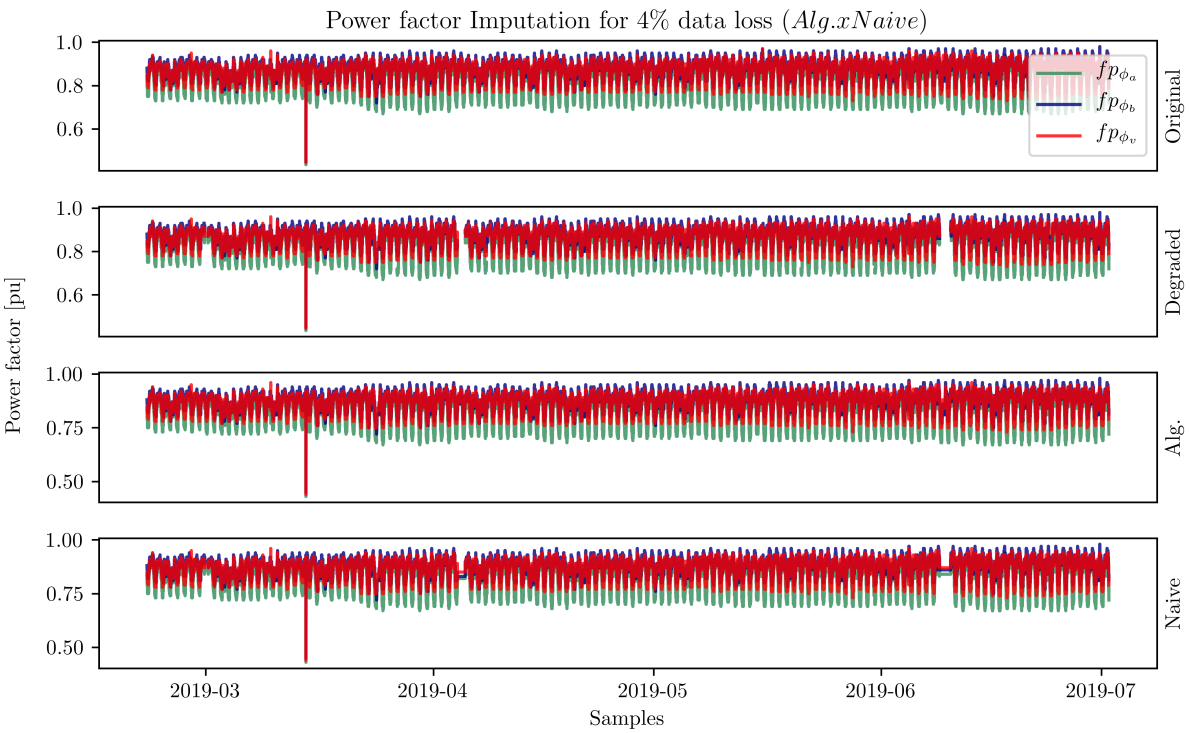

Figure 71: Example of 4% data loss on the power factor time series.

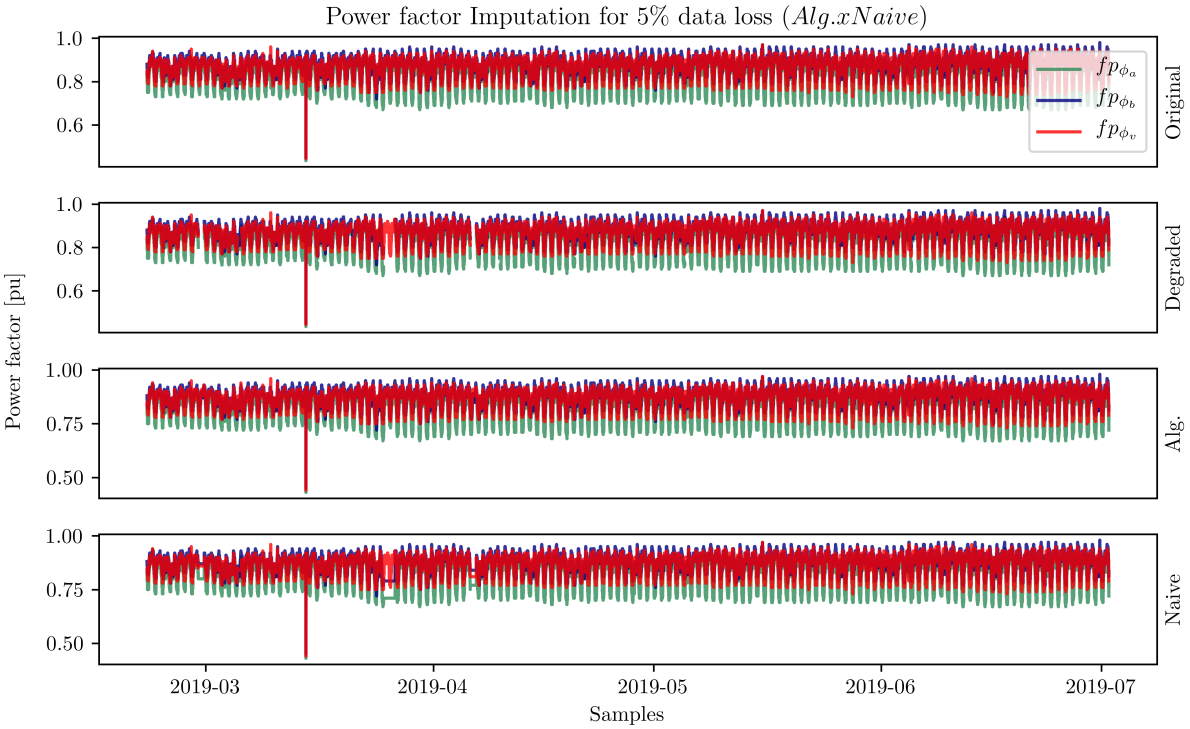

Figure 72: Example of 5% data loss on the power factor time series.

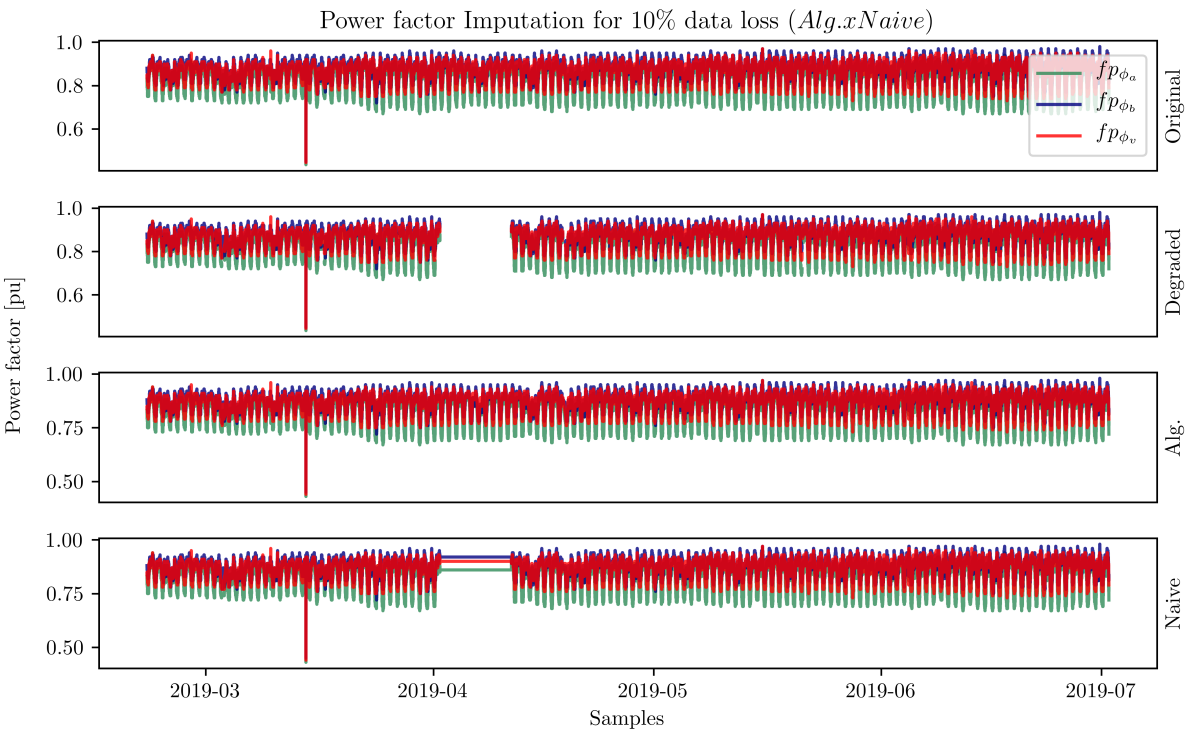

Figure 73: Example of 10% data loss on the power factor time series.

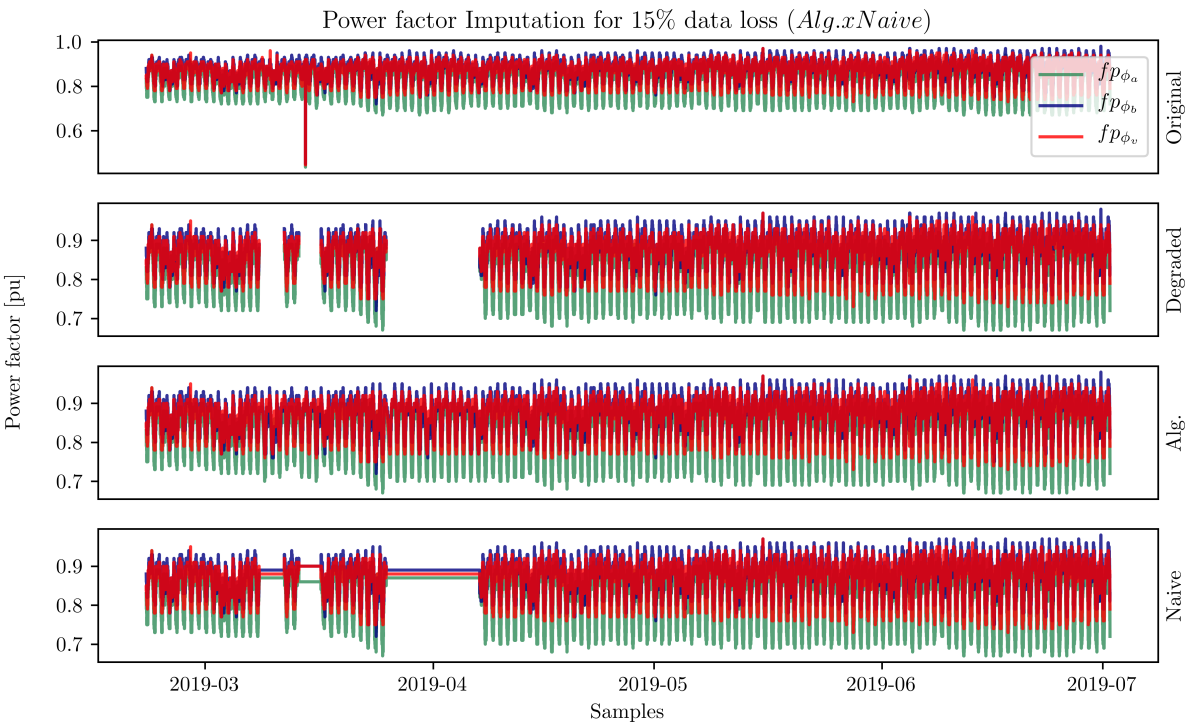

Figure 74: Example of 15% data loss on the power factor time series.

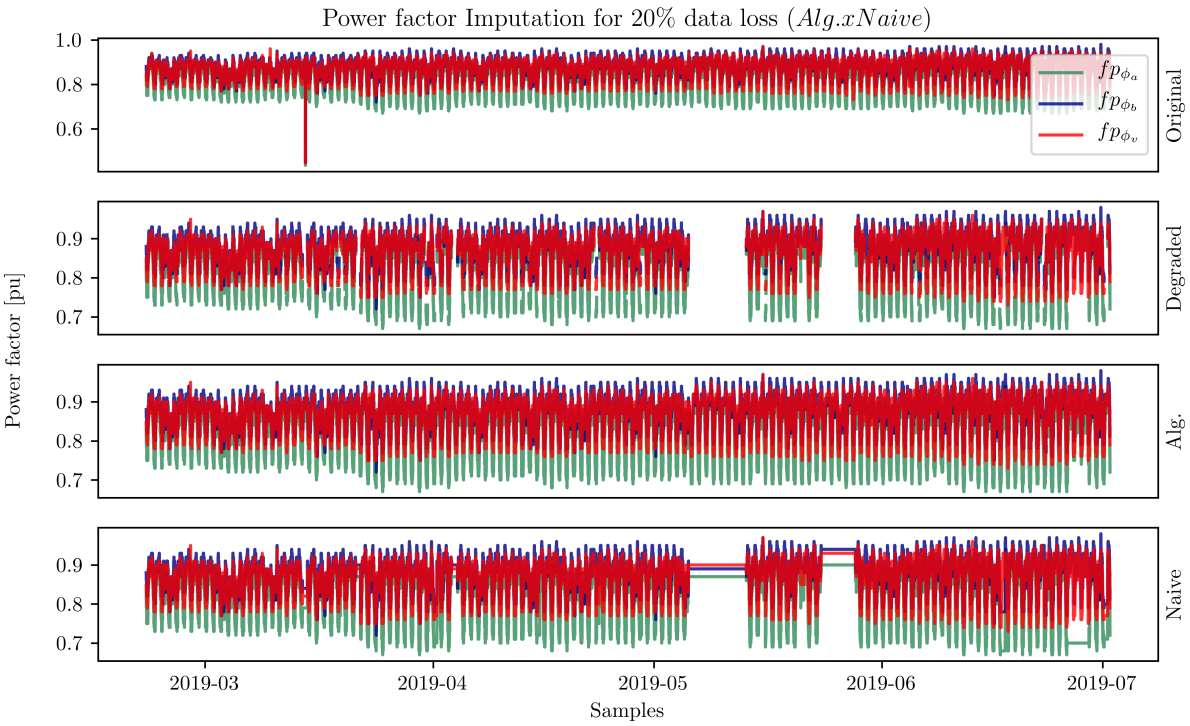

Figure 75: Example of 20% data loss on the power factor time series.

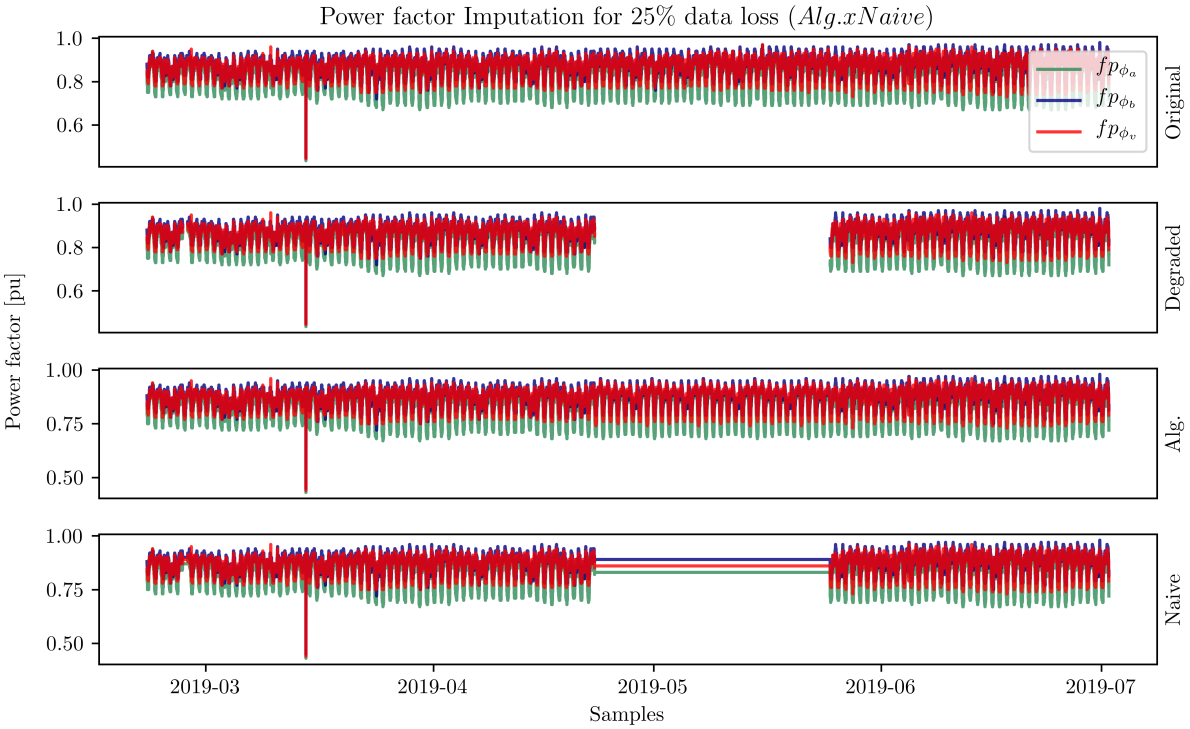

Figure 76: Example of 25% data loss on the power factor time series.

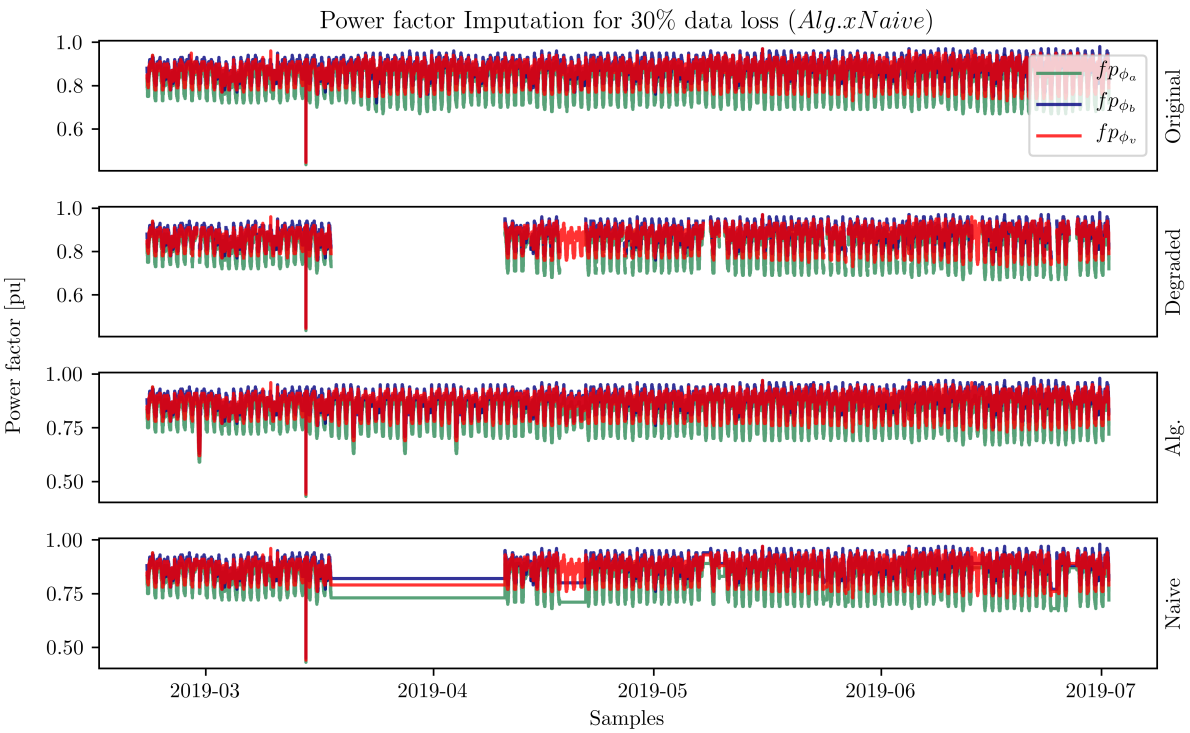

Figure 77: Example of 30% data loss on the power factor time series.

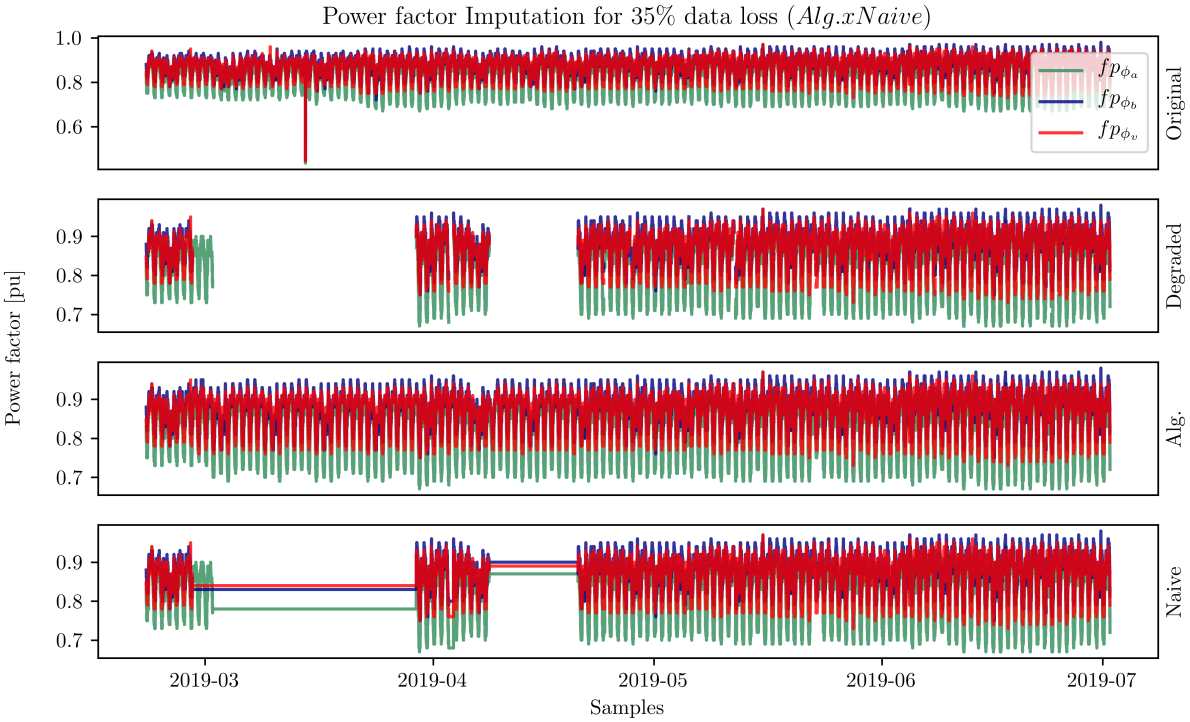

Figure 78: Example of 35% data loss on the power factor time series.

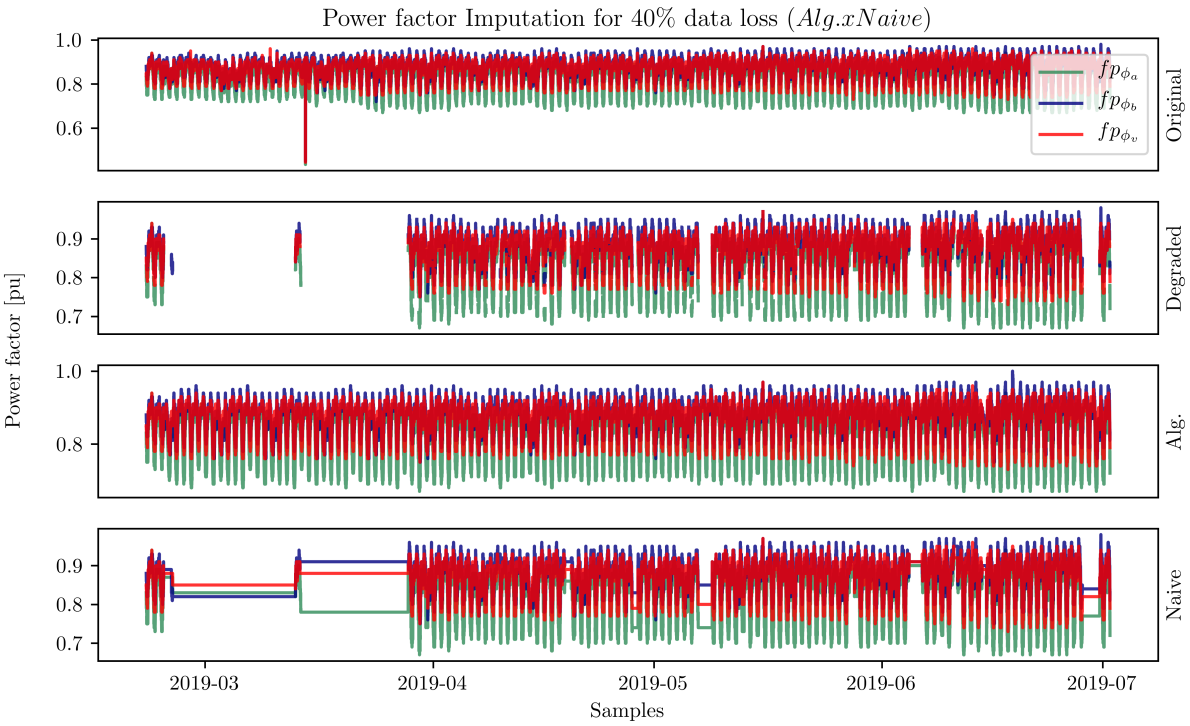

Figure 79: Example of 40% data loss on the power factor time series.

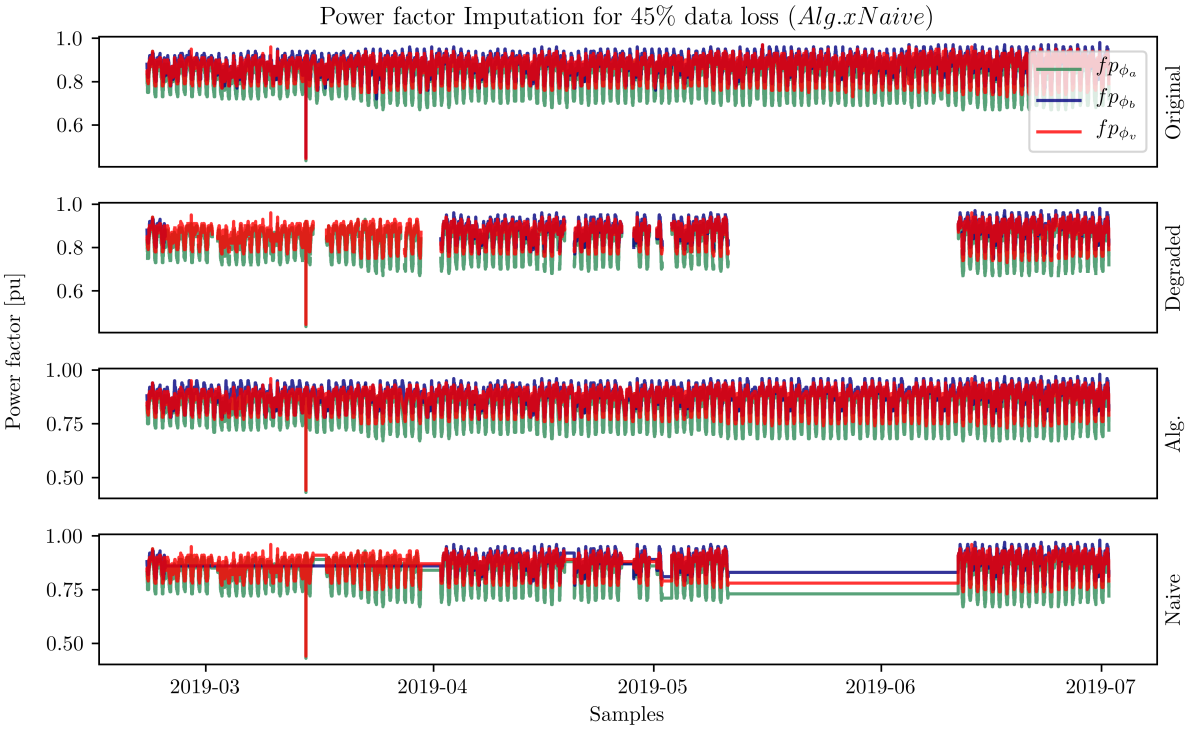

Figure 80: Example of 45% data loss on the power factor time series.

Figure 81: Example of 50% data loss on the power factor time series.

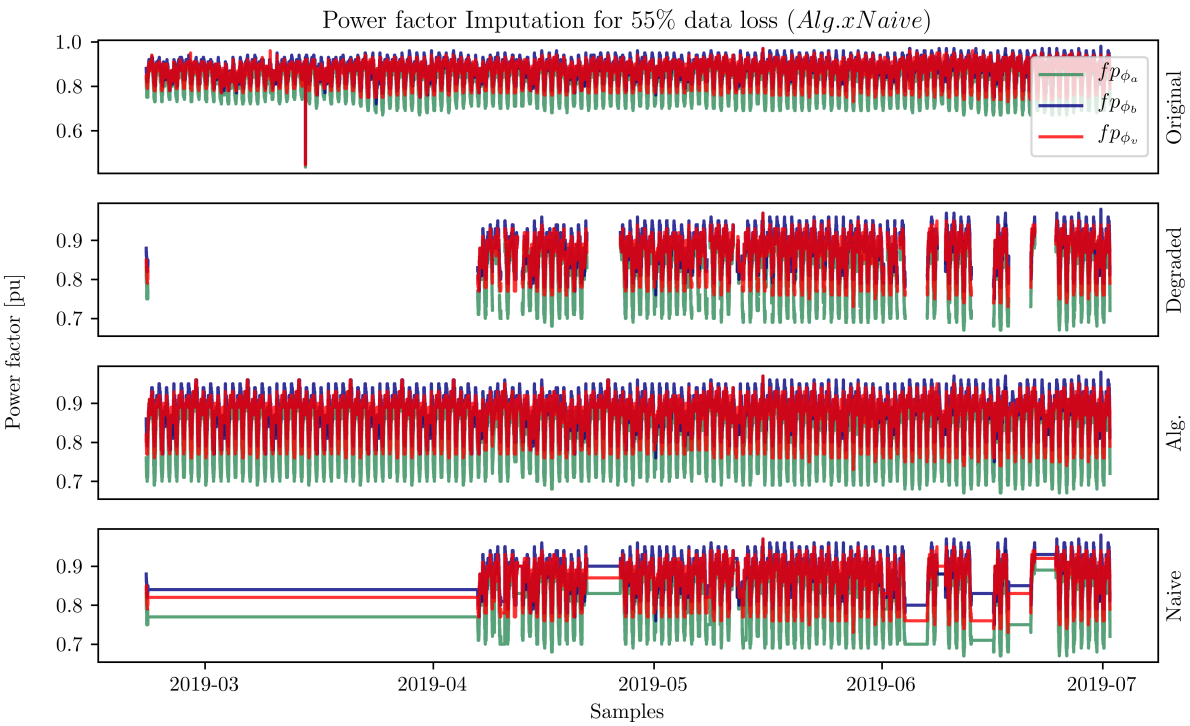

Figure 82: Example of 55% data loss on the power factor time series.

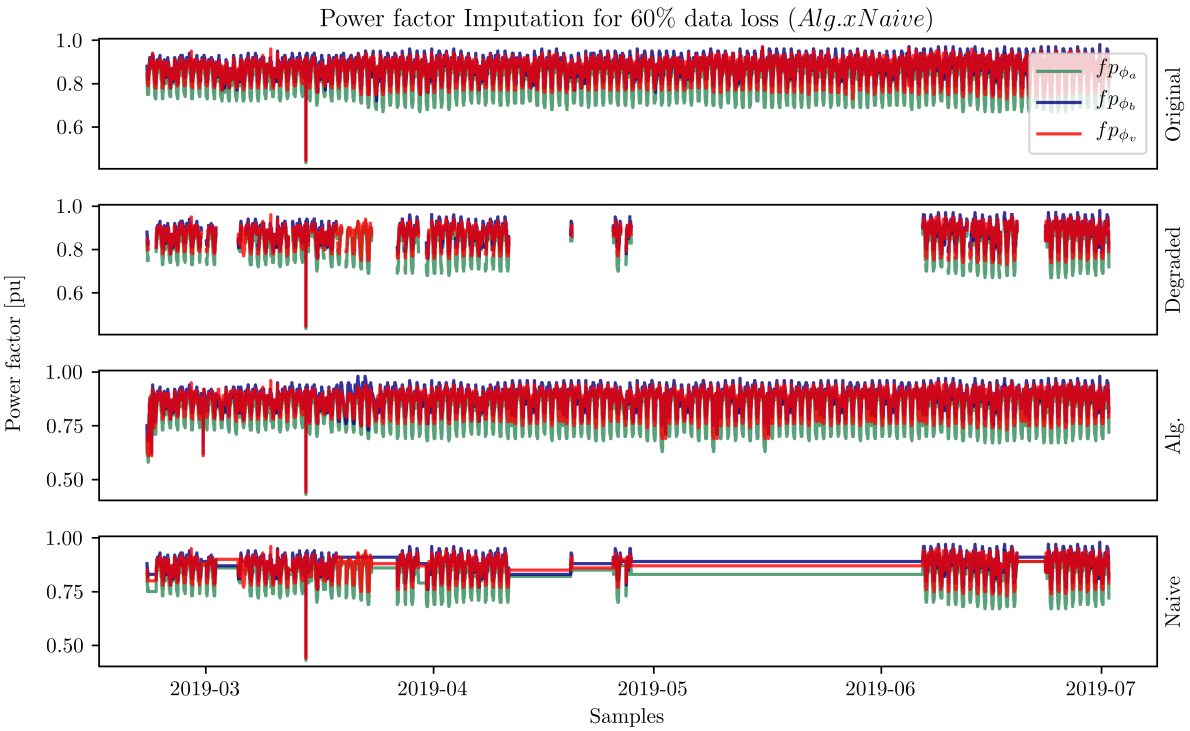

Figure 83: Example of 60% data loss on the power factor time series.

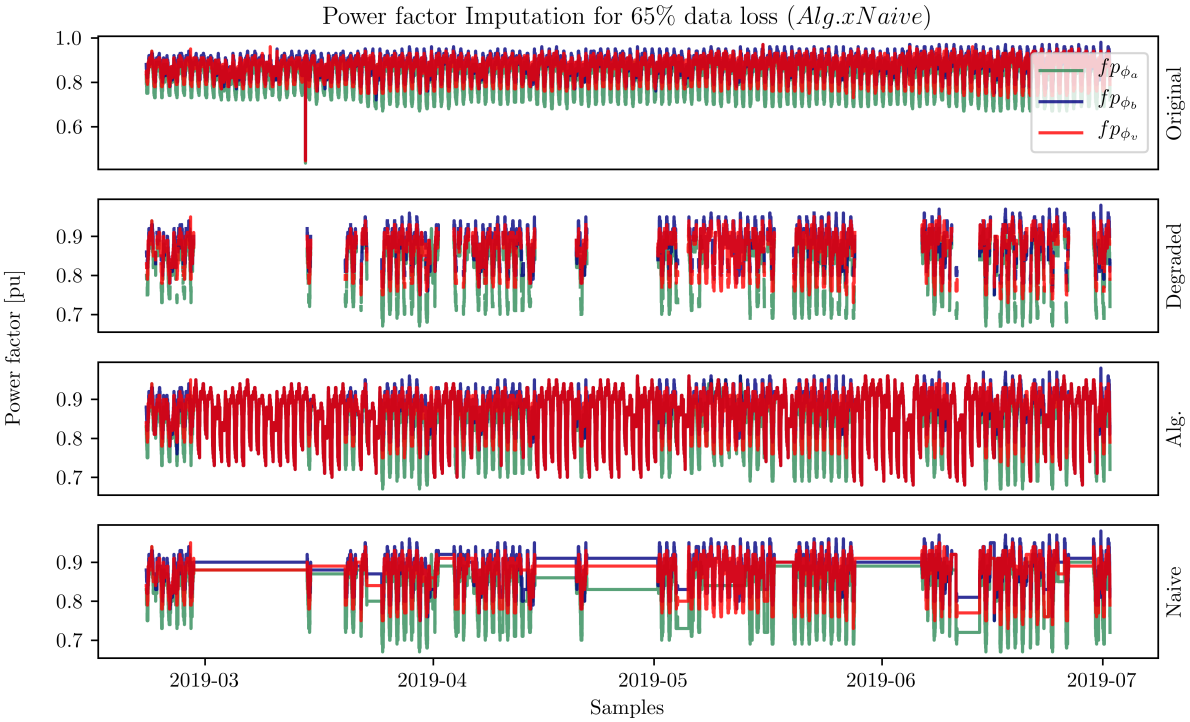

Figure 84: Example of 65% data loss on the power factor time series.

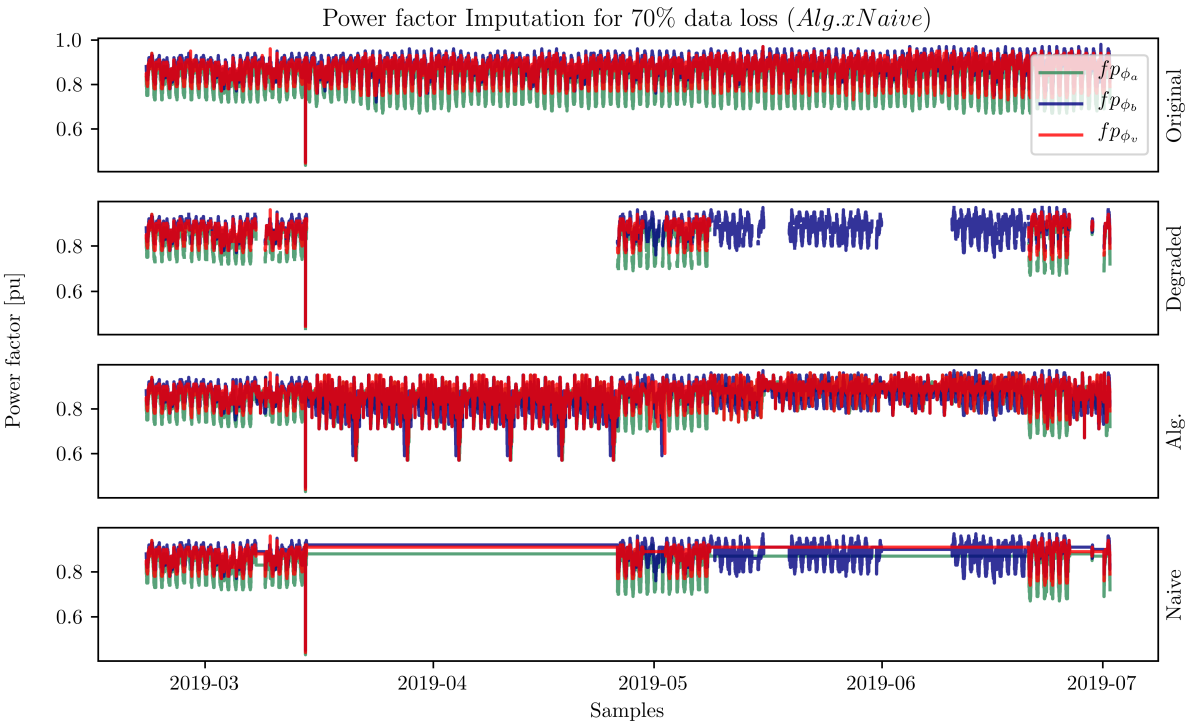

Figure 85: Example of 70% data loss on the power factor time series.

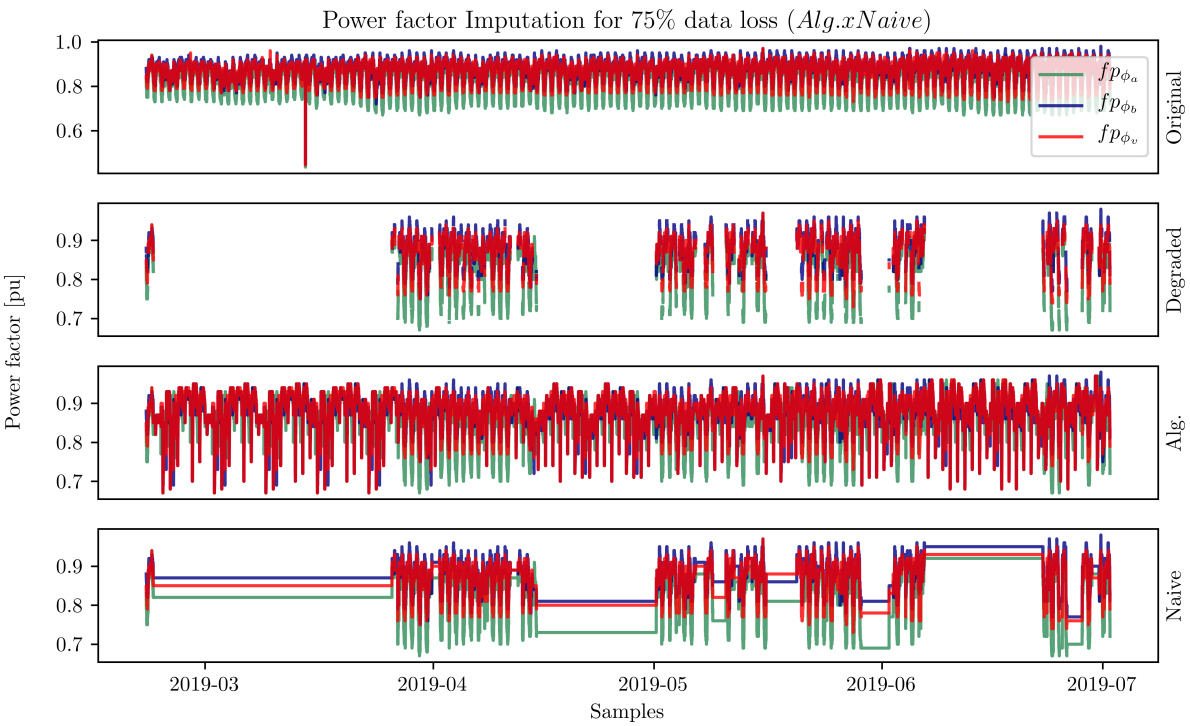

Figure 86: Example of 75% data loss on the power factor time series.

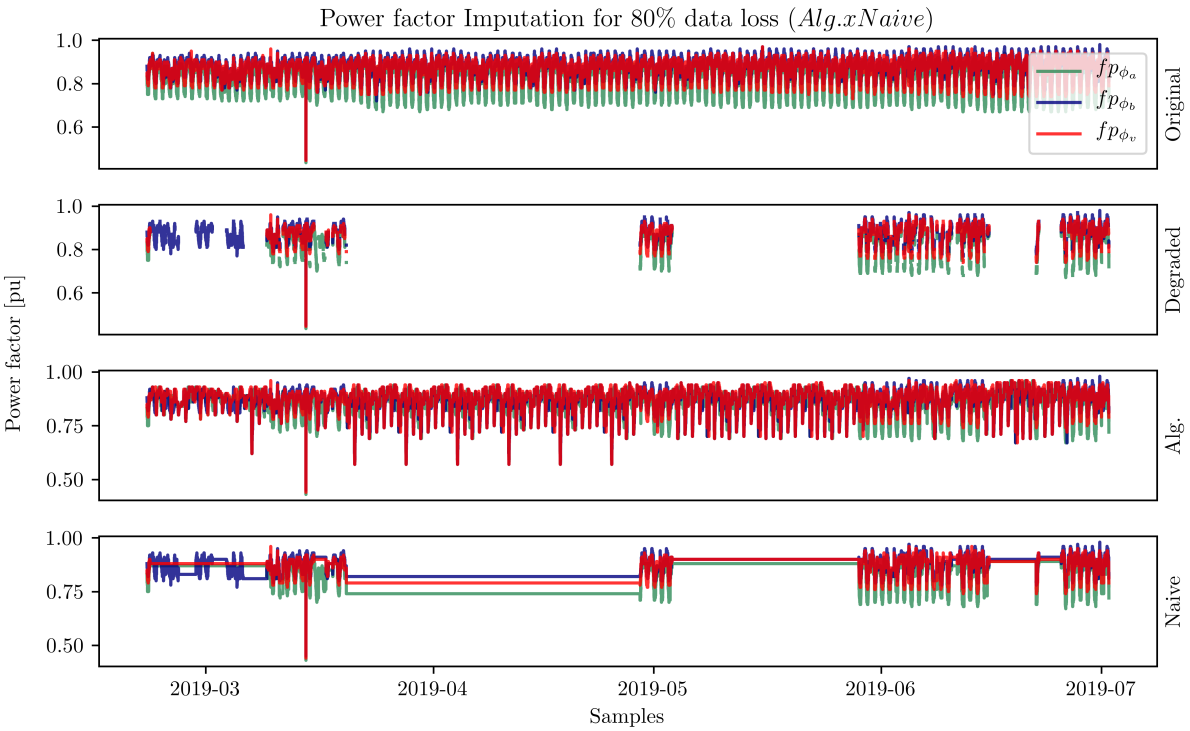

Figure 87: Example of 80% data loss on the power factor time series.

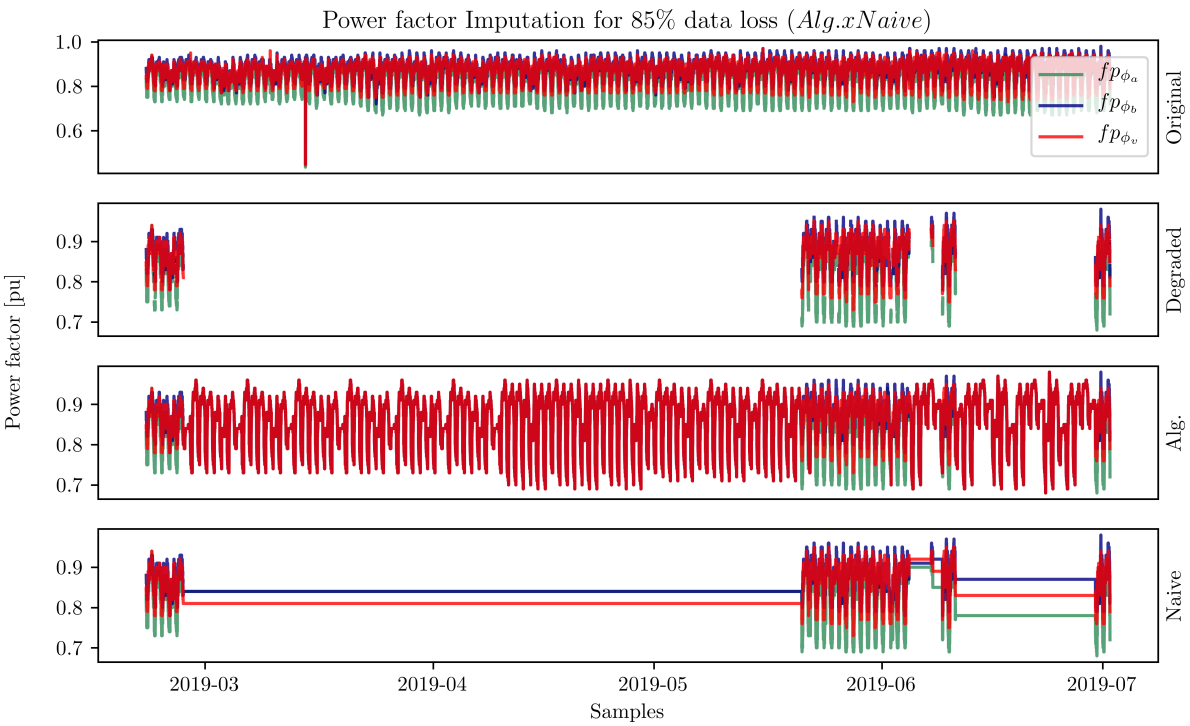

Figure 88: Example of 85% data loss on the power factor time series.

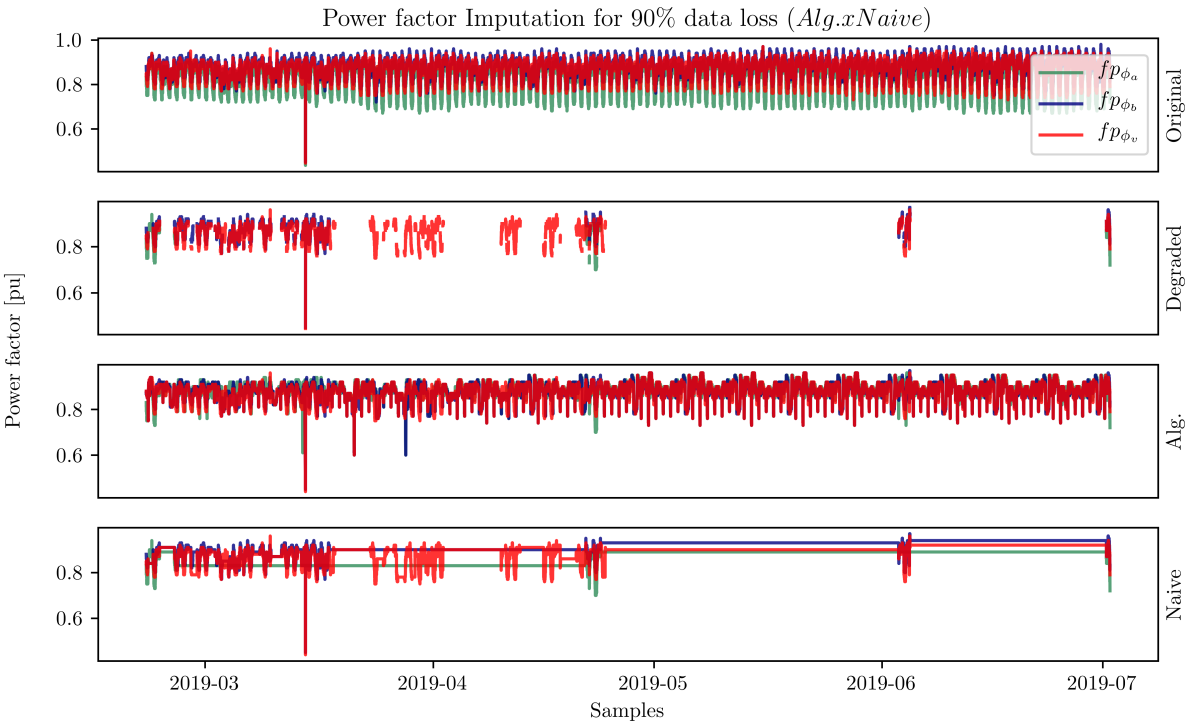

Figure 89: Example of 90% data loss on the power factor time series.

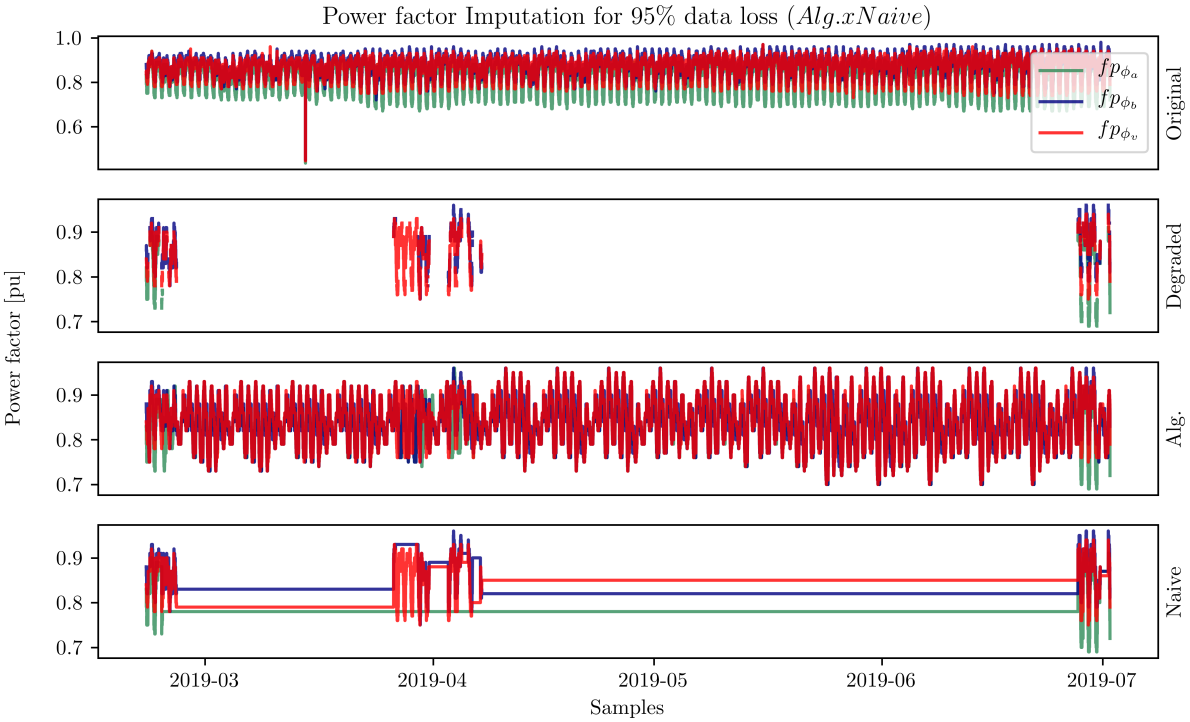

Figure 90: Example of 95% data loss on the power factor time series.