IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 July 2024, accepted 23 July 2024, date of publication 29 July 2024, date of current version 6 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3434652

RESEARCH ARTICLE

Incorporating Seasonal Features in Data Imputation Methods for Power Demand Time Series

DMITRII VASENIN[1], (Graduate Student Member, IEEE), MARCO PASETTI[1], (Member, IEEE), DAVIDE ASTOLFI[1], (Member, IEEE), NIKITA SAVVIN[2,3], STEFANO RINALDI[1], (Senior Member, IEEE), AND ALBERTO BERIZZI[4], (Senior Member, IEEE)

[1]Department of Information Engineering, Università degli Studi di Brescia, 25121 Brescia, Italy
[2]Institute for Statistical Studies and Economics of Knowledge, HSE University, 101000 Moscow, Russia
[3]Department of Automated and Computing Systems, Voronezh State Technical University, 394000 Voronezh, Russia
[4]Energy Department, Politecnico di Milano, 20156 Milan, Italy

Corresponding author: Davide Astolfi (davide.astolfi@unibs.it)

ABSTRACT This paper addresses the critical issue of missing data in power demand time series by emphasizing the relevance of imputation-based approaches in data-driven technologies. A comparative analysis of imputation methods is performed, where the reference from the state of the art is selected as K-Nearest Neighbors (KNN) applied in the time domain. Two innovative methods are proposed. The former method is defined as Historical Data Informed Regression Technique (H-DIRT) and is based on incorporating historical data for setting up a multivariate linear regression and then imputing through the estimated relation between the missing power demand measurement and the historical data. When the available historical data are insufficient, the algorithm proceeds by averaging or by a linear interpolation between the first available measurement before and after the missing value. The latter proposed method is defined as Seasonal KNN (SKNN) and it is based on enriching the data set with features related to yearly, seasonal, weekly and daily trends and then proceeding by baseline KNN. Experiments are set up with random and continuous data clipping, even with rather extreme pruning (up 70% of the data). The results in general demonstrate a significant improvement in imputation accuracy compared to the state of the art. The average error metrics (like Mean Absolute Error and Root Mean Square Error) for the SKNN method are in the order of respectively one third and one half those of the baseline KNN, in the cases of random and continuous data clipping. In general, the SKNN method provides more accurate results and better captures the statistical features of the data set to impute. Anyway, if the share of data to impute is not too large, the H-DIRT method provides comparable accuracy at a much lower computational cost. Hence, this study presents an easily implementable and computationally affordable approach for improving, in various contexts, the state of the art in power demand data imputation. It establishes a foundation for future exploration into trends, seasonal factors, and external variables influencing power load parameters.

INDEX TERMS Data imputation, electricity consumption, load analysis, data analysis.

I. INTRODUCTION

Residential and commercial buildings stand as pivotal contributors to the overall landscape of energy consumption [1]. The advent of sensor technologies and the integration of smart systems have intensified the need for precise short-term electricity consumption predictions within these structures [2]. As the energy sector seeks to navigate this era of heightened data availability, a critical challenge arises in ensuring accurate forecasting when faced with limited access to detailed building information [3], [4]. This limitation amplifies the attractiveness of data-driven Machine Learning

The associate editor coordinating the review of this manuscript and approving it for publication was Zhengmao Li.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

(ML) models [5], which can adeptly navigate uncertainties and hidden patterns in energy consumption data.

Furthermore, the demand for accurate predictions in the short-term becomes increasingly crucial in guiding effective energy management strategies [6]. Decision-makers and stakeholders rely on such forecasts to optimize resource allocation, plan infrastructure development, and implement sustainability initiatives. The intricate interplay between technological advancements, evolving consumer behaviors, and dynamic external factors necessitates a deep understanding of electricity consumption trends.

In this context, the importance of data-driven machine learning models has been growing due to their capability to glean insights from available data, despite the challenge of missing segments in the energy consumption dataset. The inherent complexities of time-series predictions are increased by the sporadic nature of these data gaps [7]. Consequently, the importance of developing robust strategies for load forecasting becomes paramount, not only for accurate predictions but also for the strategic and efficient utilization of resources in the ever-evolving energy landscape.

As the energy sector continues its trajectory towards increased reliance on data and technology, the significance of accurate load forecasting and load assessment methodologies cannot be overstated. This research endeavors to bridge the gap between the inherent challenges posed by incomplete data and the pressing need for accurate predictions, thereby offering a valuable contribution to the ongoing advancement of energy management practices.

A. AIM AND MAJOR CONTRIBUTIONS

The aim of our study is to develop computationally affordable and accurate power demand time series data imputation methods, which in particular can be helpful as a preprocessing stage in the context of short-term load forecasting or load analysis.

The major contribution of our study is the formulation of two innovative data imputation methods, which are defined as Historical Data Informed Regression Technique (H-DIRT) and Seasonal KNN (SKNN):

- H-DIRT is context-adaptive, in the sense that it integrates seasonal trends into a multivariate linear regression by employing the available measurements and imputes depending on the context. If the requested historical data are available, it imputes using the model, while, if the available historical data are insufficient, it proceeds by historical data averaging or even by linear interpolation between the first available measurement before and after the missing measurement.
- SKNN is an application of the baseline KNN upon feature enrichment by considering information on a yearly basis (holiday or working day), seasonal (number of the week in the year), weekly (number of the day in the week), daily (hour and minutes) trends;

It is notable that the proposed methods both employ additional information with respect to solely the power demand time series, but they do it in a conceptually different way. The H-DIRT employs the information related to seasonality to arrange the historical power demand measurements into a vector of features for a multivariate linear model and then the formulated model is based solely on such power demand measurements. The SKNN instead employs the information about yearly, seasonal, weekly and daily trend directly as additional features which are included in the data set. The issue of feature enrichment is crucial not only in the phase of data imputation, but also in general for the forecasting [8] and the load analysis.

A qualifying point of the present paper is that a real-world test case is considered for the application of the proposed method. Power demand measurements have been obtained from the Energy Management System (EMS) records of an electric power station situated on the engineering campus of the University of Brescia, Italy. Several years of measurements are analyzed in the present work.

In the work the following steps are conducted:

- Comprehensive analysis and comparison of the proposed methods with existing techniques (KNN is selected as reference) in terms of accuracy and robustness;
- Detailed discussion of various experiments with random and continuous data clipping from real-world data sets;
- Discussion of the potential implications of our findings on short-term load forecasting practices.

The obtained accuracy is in general much better compared to the state of the art and, in particular, it depends on the context of missing power demand measurements, since the two proposed methods incorporate the seasonal information in different ways. It is remarkable that, with both the proposed methods, it is possible to perform data imputation even if there is a large share of continuously missing data. The accuracy of the SKNN method is in general higher, which implies that the feature enrichment and data set arrangement proposed in this paper are indeed effective for successfully applying the principles of the KNN method in the time domain even in particularly disadvantageous contexts (high shares of missing data). If the share of missing data is not too high, the H-DIRT method provides results which are comparable to those of SKNN, but at a much lower computational cost. Hence, the rationale of this work is to propose two different methods, which are both more accurate than the state of the art and can adaptively be employed, with a reasonable computational cost, depending on the context.

B. ARTICLE ORGANIZATION

The analysis of existing literature pertaining to the topic is detailed in Section II. Section III briefly summarizes the state of the art, including KNN and the linear regression (employed as building block of the H-DIRT method). The proposed methods (H-DIRT and SKNN) are presented in detail in Section IV. Section V outlines the case study, the analyzed data sets, the setting of the experiments and the metrics employed for the assessment. Subsequently, Section VI presents the findings of the experiments. Finally, Section VII summarizes the outcomes of the paper.

II. LITERATURE REVIEW

Significant strides have been made in addressing missing data by mixing statistical methods and machine learning. This section reviews recent advancements in imputation techniques, focusing on machine learning, artificial intelligence, and statistical methods.

A. STATISTICAL METHODS

Despite the plethora of imputation algorithms available, traditional statistical methods are still indispensable tools for handling missing data in energy consumption datasets due to their ease of implementation, high accuracy, and acceptable computational speed. Statistical approaches encompass a wide range of techniques, starting from simple methods such as averaging [9], mode [10], or median [11], and extending to more complex techniques like principal component analysis [12] and singular value decomposition [13]. These advanced techniques involve projecting the data into a lower dimensional space, where the most important features are expected to be highlighted. The mixture factor analysis is another technique which effectively leverages daily load patterns [14].

However, the effectiveness of statistical methods diminishes in complex datasets with non-linear dynamics. Linear statistical interpolation [15] and methods using the Rosenblatt transformation [16] are common, but may produce less informative replacement values. Studies like [17] evaluate listwise deletion, predictive mean matching, and Poisson imputation, suggesting their suitability based on data distribution.

For short-term load forecasting, regression-based methods like linear regression [18] and multiple regression [19] exhibit notable accuracy and ease of implementation across diverse datasets. The simplicity and versatility of these models, including seamless integration of external parameters, can make them promising candidates for practical applications. In [20] linear regression even outperformed ML techniques, and shown better accuracy when there is a linear relationship between the target and dependent variables.

Conversely, logistic regression or multinomial logistic regression is commonly utilized for estimating missing values pertaining to categorical attributes, as these methods are well-suited to modeling categorical outcomes [21].

B. MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE METHODS

Smart grids heavily rely on smart meter data, yet missing values can prevent a correct behavior of applications. In this context, there are several Machine Learning applications for imputing missing data.

The Copy-Paste imputation method, ensuring total energy preservation, emerged as highly accurate in real-world applications [22]. Another approach, least squares support vector machine, addresses missing power measurements, outperforming utility best practices [23]. In [24], a recurrent neural network-based denoising autoencoder algorithm is proposed for gap imputation in related multivariate time series, with an application to commercial buildings data sets. A novel KNN-based strategy, as highlighted in [25], exhibits promising potential in effectively imputing missing data. Furthermore, in [26] another method for imputation is proposed, which is an elastic net regularized local least squares-based approach built upon KNN. In another study, a modified KNN method was proposed [27], where a new missing value imputation method based on denoising autoencoders with kNN for the pre-imputation task was successfully implemented. Additionally, [28] demonstrates favorable outcomes through the implementation of a KNN simple regressor to address data gaps in power-related datasets.

In [29], the authors undertook experiments on two publicly accessible datasets to assess the efficacy of Generative Adversarial Imputation Nets (GAIN), an algorithm for missing data imputation. The study compared the performance of GAIN against established state-of-the-art methodologies. The findings illustrated the superiority of the GAN-based algorithm over its counterparts, particularly evident in its lower RMSE (Root Mean Square Error) and superior Fréchet inception distance scores. In [30], it is argued that the GAN algorithm is particularly advantegeous when dealing with datasets with a high share of missing data. Furthermore, a variant called Semi-GAN [31] showed excellent results compared to traditional attribution methods when all missing data ratios in the experiments were less than 20%.

Finally, it should be noticed that studies based on various imputation methods, encompassing least squares support vector machine [32], autoregressive integrated moving average [33], and Artifical Intelligence (AI) techniques in general [34], highlight the need for adaptable and context-specific approaches.

C. SELECTION OF THE STATE OF THE ART METHODS

From Sections II-A and II-B, it can easily be argued that there is a variety of methods employed for data imputation applications. Despite the growing complexity of the recently proposed methods, some classic methods still stand out for their ease of implementation, interpretability and affordable computational cost. Linear regression, in particular, is recognized for its speed and efficiency in implementation. Moreover, the KNN method demonstrates effectiveness, simplicity in implementation, and potential for improvement through the inclusion of external features [35].

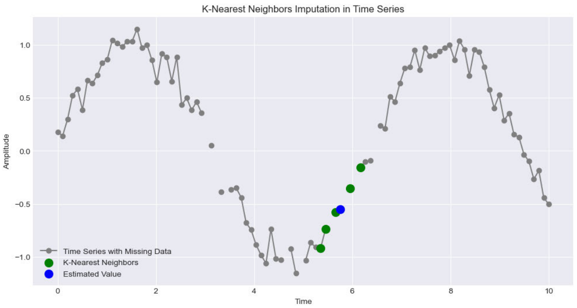

FIGURE 1. A graphical representation of a simple KNN imputation.

The computational advantage enhances the practical appeal of linear regression, making it a swift and reliable choice for handling missing data, especially in the context of complex energy consumption datasets. Similarly, the simplicity of KNN and its potential for enhancement with external features contribute to its attractiveness for addressing missing data challenges.

III. STATE OF THE ART METHODS

Based on the line of reasoning in Section II-C, linear regression and KNN are selected in this work as state of the art methods, which in the proposed methods are opportunely adapted by incorporating season-related information. Since the linear regression requires a set up and the individuation of independent variables on which the output linearly depends, for the sake of comparison against the methods proposed in this work, the KNN imputation technique [36] is selected. This is due to the fact that its application is straightforward.

Anyway, one of the building blocks of the proposed H-DIRT method is the linear regression and thus a brief outline is reported as well. Since these methods are well known, the focus is on highlighting their pros and cons in addressing data gaps in time series, particularly in load analysis.

A. K-NEAREST NEIGHBORS (KNN)

K-Nearest Neighbors (KNN) is a data imputation technique used to fill in missing values within a dataset. Given a dataset with missing values, the KNN imputation estimates them by leveraging the local structure of the data. The technique operates based on a specified parameter, K, which denotes the number of nearest neighbors considered during the imputation process. These nearest neighbors, determined using a distance metric, contribute to the estimation of the missing values (Figure 1).

- **Pros**: The main advantage of KNN is that it employs only the K nearest available measurements, which do not even need to be consecutive. This is particularly advantageous in case the available data set is not much populated. Furthermore, the KNN imputation proceeds through a weighted average and does not establish complex data-driven models.

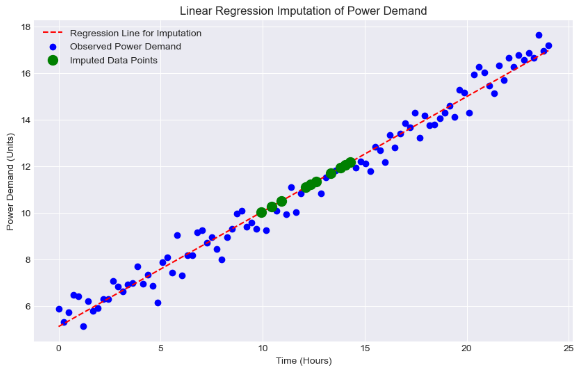

FIGURE 2. A graphical representation of the LR imputation.

- **Cons**: Since the employed information is little, the drawback of baseline KNN is that the complex behavior of the load time series might not be captured effectively.

B. LINEAR REGRESSION (LR)

In the Linear Regression (LR) data imputation technique [37], one or more independent variables (X) are selected and are supposed to be in a linear relation with a dependent variable (Y), which in this case evidently is the power demand. A representation of the LR-based imputation is provided in Figure 2, where it is supposed that the independent variable is the time and the dependent variable is the load.

- **Pros**: The advantage of the Linear Regression imputation technique is that it leverages the information contained in the historical data and thus it might be more powerful in capturing the trends in the power demand time series.
- **Cons**: The Linear Regression imputation technique is in general more demanding. It requires to establish what the independent variables are, it requires an adequate population of historical data to train the model.

IV. PROPOSED METHODS

While KNN and LR are well-established methods with different approaches but proven effectiveness, they may not fully capture the complexities of time-dependent data patterns observed in load analysis. To address this limitation, the present research introduces two novel techniques: H-DIRT and SKNN.

H-DIRT employs the historical data by setting a season-informed multivariate linear regression. For the imputation, the method evaluates the context (availability or not of the measurements requested by the linear model) and proceeds by employing the model, or data averaging, or linear interpolation between the first available measurements before and after the missing value.

On the other hand, SKNN represents an improvement of the standard KNN through feature enrichment. A richer data set is constructed by adding features related to the yearly, seasonal, weekly, daily trends. By doing this, potentially the pros of standard KNN (like the simplicity and the low quantity of requested information) are kept, while the cos (mainly, the lack of context incorporation) are circumvented through the addition of the season-related features.

A. NOMENCLATURE

Let us define:

- k: the discrete-time variable used in all the algorithms;
- $P[k]$: the electrical load measured at time k;
- $\hat{P}[k]$: the imputed electrical load at time k;
- $y[k]$: the year of the k-th load measurement;
- $m[k]$: the month of the k-th load measurement;
- $w[k]$: the ISO week number of the year of the k-th load measurement;
- $d[k]$: the day of the k-th load measurement;
- $dow[k]$: the numerical representation of the day of the week (0 for Monday, 6 for Sunday) of the k-th load measurement;
- $h[k]$: the hour of the day of the k-th load measurement;
- $min[k]$: the minutes of the hour of the k-th load measurement;
- t: represents the timestamp of the missing value;
- t_1: the timestamp associated with the first available data points before the missing value;
- t_2: the timestamp associated with the first available data points after the missing value.

B. THE HISTORICAL DATA INFORMED REGRESSION TECHNIQUE

The H-DIRT algorithm is an innovative approach designed to enhance the imputation of missing values in time series data. The method integrates linear regression with context-adaptive filling strategies to address missing values in time series data.

The H-DIRT algorithm comprises several steps and a decision making in order to select the data imputation method. Therefore, in the following, the sequence of steps and the criteria are described in sequence.

1) CONSTRUCTION OF THE CONTEXTUAL VECTOR

A contextual vector $\tilde{\mathbf{P}}[k]$ is defined for each data point $P[k]$ using data from predefined time points before and after the missing value as in Equation 1:

$$\begin{aligned}\tilde{\mathbf{P}}[k] = [&P_{y[k]-3,m[k],d[k],h[k]}P_{y[k]-2,m[k],d[k],h[k]}\\ &P_{y[k]-1,m[k],d[k],h[k]}P_{y[k]+1,m[k],d[k],h[k]}\\ &P_{y[k]+2,m[k],d[k],h[k]}P_{y[k]+3,m[k],d[k],h[k]}]\end{aligned} \tag{1}$$

In a nutshell, for each data point $P[k]$, the contextual vector $\tilde{\mathbf{P}}[k]$ is constructed by picking the data at the corresponding hour, day, month and for the previous and subsequent three years (when such data are available).

In case the share of missing data is high, it is not assured that the contextual vector is composed of all valid measurements. In other words, if we want to impute the k-th measurement, there might be missing values also in $\tilde{\mathbf{P}}[k]$. As detailed in the following Section IV-B2, at this point, the steps for the data imputation are selected based on the completeness or not of the the contextual vector.

2) DECISION-MAKING FOR IMPUTATION METHOD

- **Criterion**: The choice of imputation method depends on the completeness of the contextual vector.
- **Options**:
  1) *Complete contextual vector*: Apply a multivariate linear regression.
  2) *Partially complete contextual vector*: Use averaging of the data available in the partially complete contextual vector.
  3) *Empty contextual vector*: Employ linear interpolation between the first available data points before and after the missing data.

a: LINEAR REGRESSION

- *When Applied*: When the contextual vector $\tilde{\mathbf{P}}[k]$ around a missing k-th data point is complete. This means all its values are available and non-missing.
- *Model Implementation and Training*: Suppose that there is a linear relation between the complete contextual vectors data and the power demand data. Thus the independent variables are supposed to be the data points from the contextual vector $\tilde{\mathbf{P}}[k]$ defined in Equation 1. The dependent variable is the electricity load measurement $P[k]$ at the time k. The linear regression function can be posed as in Equation 2:

$$P = \beta_0 + \tilde{\mathbf{P}} \cdot \boldsymbol{\beta} + \epsilon \tag{2}$$

  where β_0 is a the intercept and $\boldsymbol{\beta}$ is a vector of six model parameters, and ϵ is the model error. Using the Least Squares Method, the best estimate $\hat{\boldsymbol{\beta}}$ of the coefficients $\boldsymbol{\beta}$ is obtained.
- *Imputation of the missing data*. Using the trained linear regression model, the missing value can be predicted based on the values in the contextual vector $\tilde{\mathbf{P}}[k]$. The estimated value $\hat{P}[k]$ is thus calculated as in Equation 5

$$\hat{P}[k] = \hat{\beta}_0 + \tilde{\mathbf{P}}[k] \cdot \hat{\boldsymbol{\beta}}. \tag{3}$$

b: AVERAGING OF VALUES

- *When Applied*: Used when the contextual vector is partially complete, meaning at least one but not all of the six contextual values are non-missing.
- *Imputation of the missing data*. The estimate $\hat{P}[k]$ is the average of the available data points within the contextual vector $\tilde{\mathbf{P}}[k]$, which in this case is incomplete.

c: LINEAR INTERPOLATION

- *When Applied*: Employed when the contextual vector of Equation 1 does not contain valid measurements. The method assumes that there are at least two known points (before and after the missing value), but not enough context for a more sophisticated approach.

- *Imputation of the missing data*. Linear interpolation is straightforwardly applied between the first available data point before and the first available data after the missing value. Suppose that k_1 and k_2 are the indexes associated to the first available data point before and after the missing values, respectively, and the corresponding times are t_1 and t_2. This process draws a straight line between these two known points and estimates the missing value based on its proportional position along this line.

C. THE ENHANCED KNN METHOD WITH SEASONAL INPUT FRAMEWORK (SKNN)

This part provides the implementation of a data imputation technique using KNN with a seasonal component. Differently with respect to H-DIRT, SKNN does not involve criteria and decision making and thus the steps are simply listed here on.

1) DATA PREPARATION

- *Extraction of the temporal features*. For each timestamp k, the function employs the following temporal attributes which are used also by H-DIRT for the contextual vector:
  - -- $y[k]$;
  - -- $h[k]$;
  - -- $min[k]$;

  And it employs the following further temporal features:
  - -- Week: The ISO week number $w[k]$ of the year corresponding to each timestamp.
  - -- Day of the Week: The numerical representation $dow[k]$ of the day of the week (0 for Monday, 6 for Sunday).
- *Trigonometric transformations of the temporal features*. Trigonometric transformations are applied to capture cyclic patterns:
  - -- Hour Sine and Cosine: Trigonometric transformations of the hour, capturing cyclic patterns within a day.

$$\text{h\_sin}[k] = \sin\left(\frac{2\pi \cdot \text{h}[k]}{24}\right) \tag{4}$$

$$\text{h\_cos}[k] = \cos\left(\frac{2\pi \cdot \text{h}[k]}{24}\right) \tag{5}$$

  - -- Day of the Week Sine and Cosine: Trigonometric transformations of the day of the week, capturing cyclic patterns within a week.

$$\text{dow\_sin}[k] = \sin\left(\frac{2\pi \cdot \text{dow}[k]}{7}\right) \tag{6}$$

$$\text{dow\_cos}[k] = \cos\left(\frac{2\pi \cdot \text{dow}[k]}{7}\right) \tag{7}$$

  - -- Week Sine and Cosine: Trigonometric transformations of the week number, capturing cyclic patterns within a year.

$$\text{w\_sin}[k] = \sin\left(\frac{2\pi \cdot \text{w}[k]}{52}\right) \tag{8}$$

$$\text{w\_cos}[k] = \cos\left(\frac{2\pi \cdot \text{w}[k]}{52}\right) \tag{9}$$

- *Extraction of the holiday information*. The function incorporates the information about the fact that the k-th measurement belong to a working day or not in Italy. If it does, a binary flag $\text{is\_holiday}[k] = 1$ is set; otherwise, $\text{is\_holiday}[k] = 0$.
- *Resulting Data Set Arrangement*. The resulting data frame contains the original load measurement $P[k]$ along with these additional features.
  - -- $y[k]$;
  - -- $\text{w\_sin}[k]$;
  - -- $\text{w\_cos}[k]$;
  - -- $\text{h\_sin}[k]$;
  - -- $\text{h\_cos}[k]$;
  - -- $min[k]$;
  - -- $\text{dow\_sin}[k]$;
  - -- $\text{dow\_cos}[k]$;
  - -- $\text{is\_holiday}[k]$.

  Thus, for each k-th time step, the data set has been augmented by passing from a scalar $P[k]$ to a season-informed vector $\mathbf{P}_{SI}[k]$.

2) DATA IMPUTATION

Consider the input data frame where the k-th measurement is missing, which has been enriched with the above listed features. The following steps are then implemented.

- Starting from $\mathbf{P}_{SI}[k]$, by cycling on the data set of available valid load measurements, the nearest neighbors data are retrieved, where the proximity of a measurement $\mathbf{P}_{SI}[k']$ with respect to $\mathbf{P}_{SI}[k]$ is computed through the euclidean distance between the two vectors.
- Upon a sensitivity analysis, $K = 5$, indicating the number of nearest neighbors to consider, is selected.
- The imputed measurement $\hat{P}[k]$ is given by the weighted average of the load components of the K nearest neighbor above individuated vectors, where the weights are the inverse of the above computed euclidean distances.

V. EXPERIMENTAL SETUP

A. DATA DESCRIPTION

1) TYPE OF DATA

The dataset used in this study was obtained from the energy management system (EMS) records of an electric power station situated on the engineering campus of the University of Brescia, Italy [38]. This facility serves a public building housing various amenities, including department facilities such as offices and classrooms, as well as student services like dormitories, study halls, a cafeteria, a gym, and a baseball field [39].

Data collection was conducted at a sampling time interval of 2 seconds. Subsequently, the recorded data were stored in a time series database, where each entry represents the average of sampled values over a 5-minute time interval and covers the period between July 10, 2016 and September 11, 2023. The averaging time has been selected because it is the typical one in load forecast or assessment applications [40], [41].

The schematic diagram on Fig. 3 illustrates the architecture of the data collection system. Components include PV (Photovoltaic), BESS (Battery Energy Storage System), BMS (Battery Management System), BM (Battery Module), BMU (Battery Management Unit), PCS (Power Conversion System), PCC (Point of Common Coupling), AC (Alternating Current), DC (Direct Current), and SOC (State of Charge).

The measurements of load consumption (P) are extracted from five different features, as shown on Fig. 4, and calculated as following:

$$P = P_{\mathrm{PV}} + P_{\mathrm{FG}} + P_{\mathrm{FS}} - (P_{\mathrm{TG}} + P_{\mathrm{TS}}) \tag{10}$$

which are power injected To the Grid (P_{TG}), power consumed From the Grid (P_{FG}), AC Power injected To Storage (P_{TS}), AC Power consumed From Storage (P_{FS}), and PV Production (P_{PV}).

In the dataset, approximately 18.51% of the rows have missing data, where (given Equation 10) it is intended that at least one among P_{TG}, P_{FG}, P_{TS}, P_{FS}, and P_{PV} is missing. The longest sequence of continuous missing values is 47589 rows long, that amounts approximately to 6.24% or 165 days of the total dataset. The representation of such missing data sequence is shown on Fig. 4 and 5. The remaining missing data, which are not part of this continuous sequence (i.e., random missing values and short continues missing values), constitute approximately 12.27% of the dataset. In summary, the dataset has both a significant continuous missing data segment and random missing data points, with the random missing data being nearly double the continuous missing data in terms of percentage.

2) DATA COLLECTION

Data retrieval from the smart1.eu platform was executed using a custom-developed software developed by a Python language, specifically designed to interact seamlessly with the smart1.eu API. To optimize the data extraction process, tags were employed as a selective filter, pinpointing only the necessary sensors from the extensive dataset available on smart1.eu. This focused approach was crucial in extracting only relevant and essential data, thus maintaining the precision and efficiency of the entire migration process.

In preparation for migration, a comprehensive CSV file was constructed. This file served as a pivotal configuration guide, systematically listing the sensors and data types set earmarked for migration. The migration to InfluxDB (Fig. 4) was facilitated by a program, tailored for efficient data transfer, which utilized a specially generated configuration file to ensure a streamlined and error-free migration process.

To enhance understanding and provide a visual representation of this process, a detailed flowchart was developed on Fig. 6. This flowchart illustrates the entire data collection and migration journey, offering a clear, step-by-step visual guide that complements the textual descriptions. The process described as following:

- *User Interaction*;
- *Sensor Selection*;
- *Data Frame Creation*;
- *API Data Collection*;
- *Data Storage*;
- *Data Visualization*.

By following these steps (detailed in Fig. 6), the application streamlines the process of migrating data from the Smart1 API cloud, ensuring efficient data management and analysis for eLux lab's operations.

This comprehensive and systematic approach to data collection, underpinned by the use of efficient, custom-developed tools, ensures the accuracy and completeness of data migration from smart1.eu to InfluxDB. Diverging from the often-simulated data used in research, our methodology capitalizes on real data. This approach not only enriches our research with authentic, real-world insights but also significantly bolsters the reliability and applicability of our findings.

B. EXPERIMENTAL ARRANGEMENT

In this subsection, we elaborate on the experimental setup and methodologies employed to address missing data within our dataset, utilizing three distinct imputation methods: KNN, SKNN, and the H-DIRT model. The aim is to systematically compare these methods against each other to evaluate their effectiveness.

The testing set for simulating the data removal is September 2022 to December 2023, which ensures that the selected period largely lacks missing values (compare against Fig. 5), providing a robust foundation for evaluating the imputation methods.

To ensure a comprehensive evaluation, each method is applied to impute data across various degrees, allowing for a thorough assessment of their performance under different conditions. Four percentages of random missing data (namely, a_1, a_2, a_3, and a_4) correspond to evaluations conducted with missing values of 10%, 30%, 50%, and 70%, respectively. Several experiments are carried out for testing the imputation of the missing values in random and continuous data pruning, with the above percentages. This approach enables us to examine how each imputation method performs across a spectrum of challenging missing data scenarios, providing valuable insights into their robustness and reliability.

It should be specified that the experiments proceed by imputing all the share of missing data (namely, a_1, a_2, a_3, and a_4) by using solely the available valid measurements, and

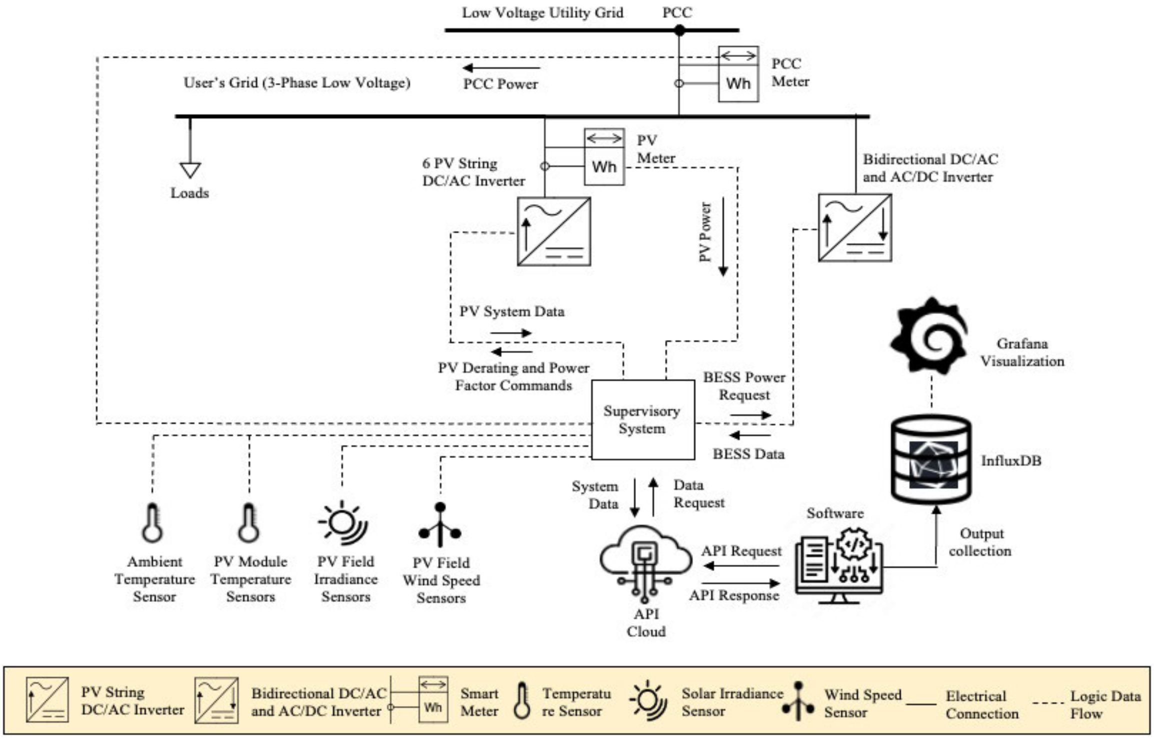

FIGURE 3. The architecture of the data collection system: Sensor and device layout.

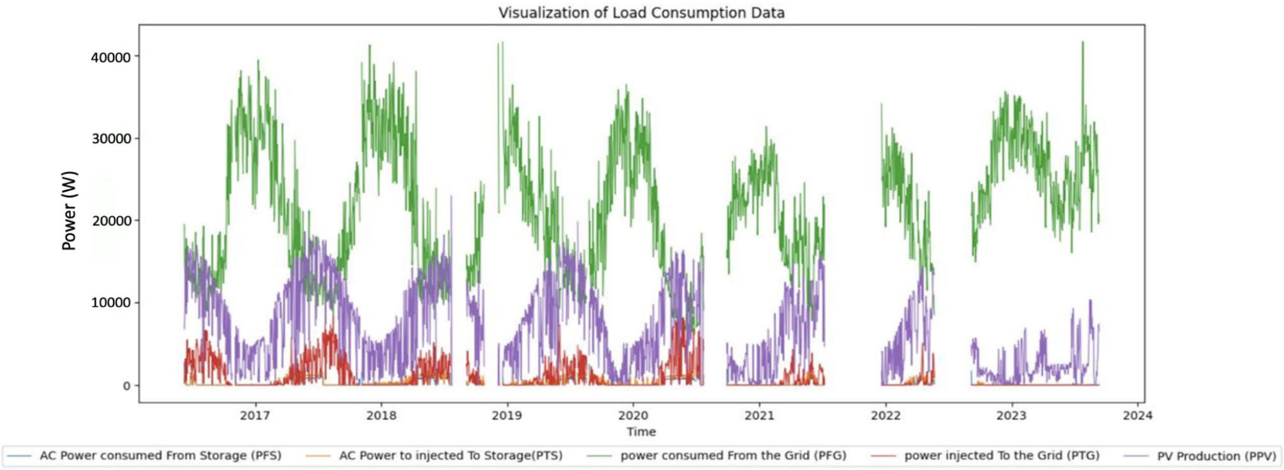

FIGURE 4. Representation of electricity consumption by five different features.

not by also leveraging on the imputed data, as long as the imputation process proceeds spanning the data set.

C. DATA IMPUTATION ASSESSMENT

1) OVERALL METRICS

The assessment of the accuracy of the various data imputation methods is based on the analysis of key metrics, providing insights into the efficacy of each imputation technique. The selected metrics are RMSE, MAE, and MAPE.

For each metric and for each couple of considered methods, a systematic comparison can be achieved by applying Equation 11:

$$\Delta = \left(\frac{\text{Method}_1 - \text{Method}_2}{\text{Method}_2} \right) \tag{11}$$

a: ROOT MEAN SQUARE ERROR (RMSE)

Root Mean Square Error is a widely used metric to measure the average deviation of imputed values from actual values. It is calculated as the square root of the average of squared

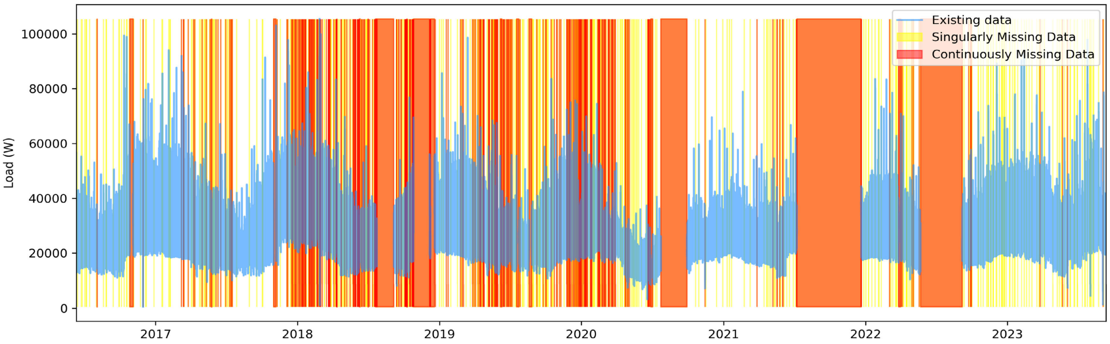

FIGURE 5. The distribution of missing data for the Ex Emiliani building dataset is depicted with the x-axis representing the time of measurements and the y-axis showing the electric load measurements. The periods during which data are missing are highlighted by red and orange shaded regions.

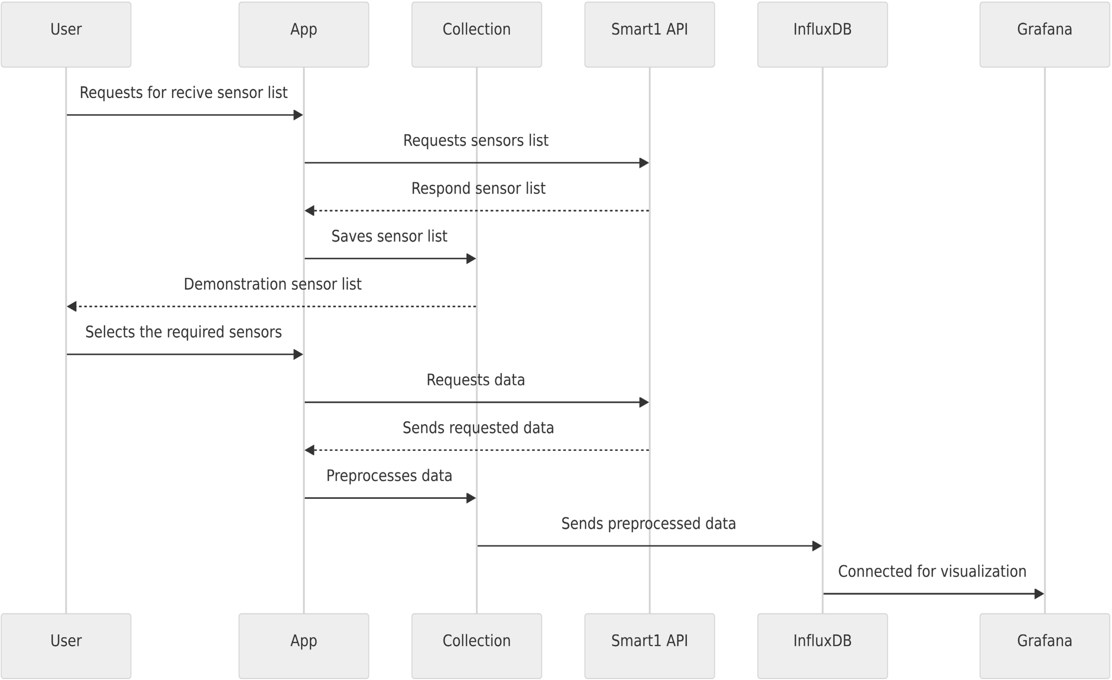

FIGURE 6. Flowchart of data collection and migration process using UML notation.

differences between imputed and actual values. The formula for RMSE is given in (12):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(P[k] - \hat{P}[k])^2} \tag{12}$$

b: MEAN ABSOLUTE ERROR (MAE)

Mean Absolute Error is a measure of the average absolute difference between actual and imputed values. It is calculated as the average of the absolute differences between the imputed and actual values. The formula for MAE is given in (13):

$$\mathrm{MAE} = \frac{1}{n}\sum_{k=1}^{n}|P[k] - \hat{P}[k]| \tag{13}$$

c: MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

Mean Absolute Percentage Error measures the average absolute percentage difference between actual and imputed values. It is often used to understand the accuracy of forecasts. The formula for MAPE is given in (14):

$$\mathrm{MAPE} = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{P[k] - \hat{P}[k]}{P[k]}\right| \tag{14}$$

2) ANALYSIS OF THE ACCURACY METRICS PER LOAD DEMAND CLASS

For the case of continuous data clipping, which is the most challenging, the metrics are computed also upon organizing the measured and imputed data in ten intervals spanning from minimum to maximum measured load. This is done in order to highlight if there are observable trends in the accuracy of

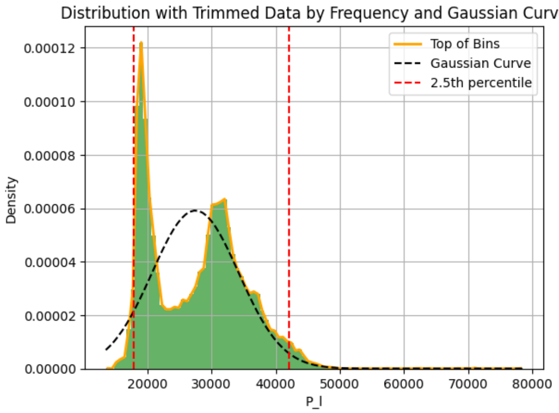

FIGURE 7. The distribution of the continuously clipped data for one experiment in the 10% case. The upper and lower 2.5th percentiles are reported.

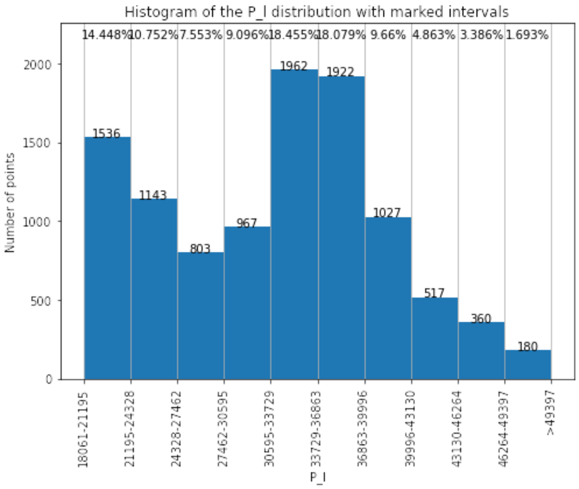

FIGURE 8. The histogram of the continuously clipped data for one experiment in the 10% case, upon filtering out the upper and lower 2.5th percentiles.

the various methods, as a function of the load demand to impute or of the share of missing data. Such analysis requires a data processing in order to guarantee a fair population to each bin and thus to make the results comparable. Thus, upon the imputation process, a pre-processing step is applied, in that the data above and below the upper and lower 2.5th percentiles are excluded. An example of this pre-processing step for one experiment arrangement is reported in Figure 7.

The resulting distribution for the various load demand bins for the same experiment arrangement is reported in Figure 8.

3) STATISTICAL ANALYSIS OF THE MEASURED AND IMPUTED DATA

In order to profitably apply the proposed methods, for example in the context of time series analysis or forecast, it is fundamental that the imputation does not alter the statistical properties of the time series. Therefore, for the case of continuous data clipping, the mean and the standard deviation of the imputed and measured data are compared. The mean and standard deviation are compared as well for the whole validation data set with data imputed or measured. Furthermore, the distributions of the real data set and of the data set with data imputed are compared.

TABLE 1. Comparison of the metrics of methods for filling in missing values in random data clipping for the best experiment.

Type	% NaN	Metric	KNN	SKNN	H-DIRT
a_1	10%	RMSE (W)	7483.1	2763.2	5948.4
		MAE (W)	5783.7	1943.8	4529.8
		MAPE (%)	21.2	6.6	16.6
a_2	30%	RMSE (W)	8510.9	3161.9	5391.2
		MAE (W)	6010.6	2139.9	4526.9
		MAPE (%)	22.1	7.2	18.1
a_3	50%	RMSE (W)	8047.5	3300.3	5898.1
		MAE (W)	6146.4	2285.5	4616.0
		MAPE (%)	24.1	7.8	18.6
a_4	70%	RMSE (W)	8303.8	3512.5	6055.8
		MAE (W)	6662.8	2459.6	4677.1
		MAPE (%)	26.6	8.4	17.7

VI. EXPERIMENTAL RESULTS

This Section presents the simulation results considering three techniques: KNN, SKNN and H-DIRT. From the dataset, the four 10%, 30% 50% and 70% of records are respectively removed from the original database in 2 different ways: continuously and randomly.

A. RANDOM DATA CLIPPING

In Table 1, we present a comprehensive comparison of the proposed H-DIRT model against state-of-the-art imputation method KNN. Each row represents a different percentage (% NaN) of missing values, and the metrics include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). In Figure 9, also the improvement provided by SKNN or H-DIRT with respect to standard KNN is reported. The takeaway message from Table 1 and Figure 9 is that in the case of random data clipping, the incorporation of the seasonal features into the KNN method, thus resulting in the SKNN method proposed in this work, provides by far the best results, with an excellent reduction of the error metrics with respect to the KNN case. Also the H-DIRT method provides a clear improvement of the error metrics for each scenario reported in Table 1, but the improvement is sensibly lower than applying SKNN for all the experiments. Hence, one of the key results of the present work is that for random data clipping, the use of the SKNN proposed in this work is preferable.

B. CONTINUOUS DATA CLIPPING

1) OVERALL METRICS

Table 2 provides a detailed comparison across three experiments, each featuring different percentages (% NaN) of missing values. The experiments were designed by randomly excising continuous segments from the dataset, ensuring that

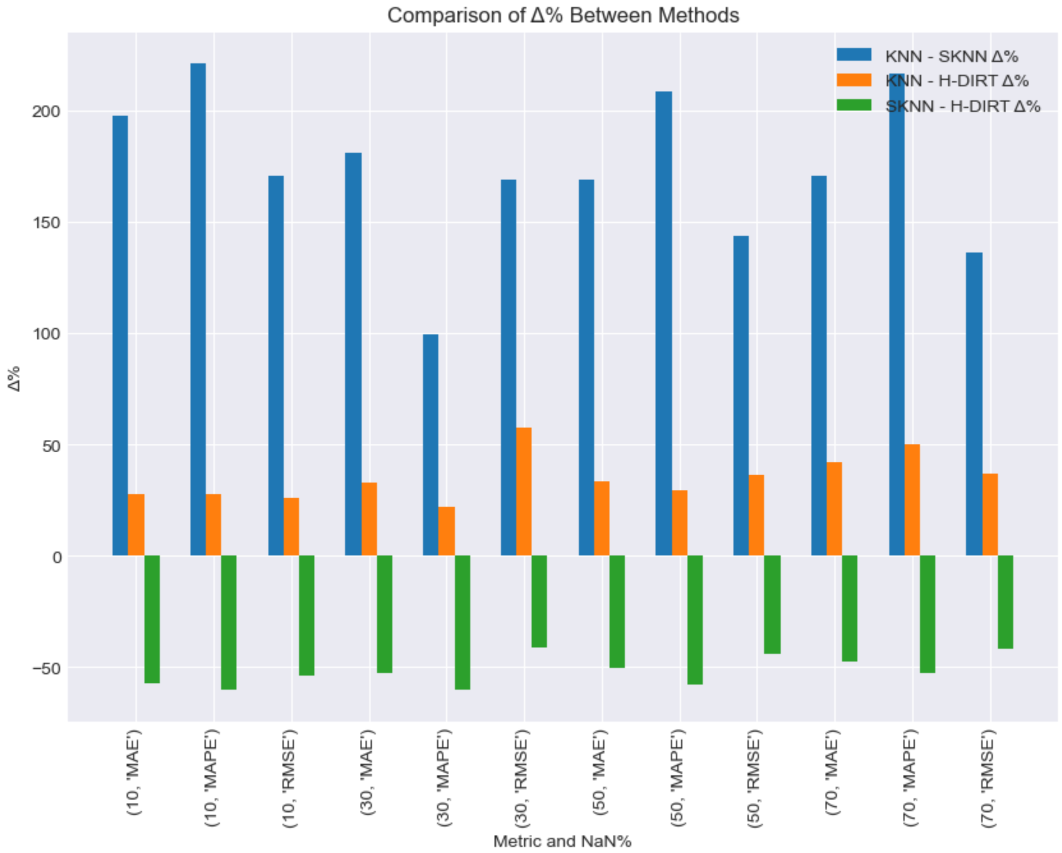

FIGURE 9. The percentage variations in the accuracy metrics for the proposed methods compared to baseline KNN and between themselves.

the removed sections could occur at any point within the data—beginning, middle, or end. This method of random selection was employed to mimic realistic scenarios where data might be missing due to various unpredictable factors. The segments chosen for detailed analysis in this table represent the top three outcomes in terms of average performance of all the methods from a series of trials, and importantly, these segments were located in different parts of the dataset, hence enhancing the representativeness of the results. This selection was made to highlight the variability in the results and to showcase the best potential outcomes under diverse scenarios of data absence. Metrics for each method are reported for these three chosen experiments to provide a comprehensive view of performance variability, crucial in the context of continuous data clipping due to the significant proportion of data removed.

The common ground of all the experiments reported in Table 2 is that there is a difference in the behavior of SKNN and H-DIRT as a function of the fraction of data clipping. Actually, for 10% of data clipping the two methods have similar error metrics, sensibly lower than the baseline KNN. As the percentage of data clipping increases, on the one hand SKNN keeps performing remarkably better than KNN, while H-DIRT does not. Indeed, in the extreme cases of 50% and 70% data clipping, it becomes less unfavorable to predict that the missing data are all constant and imputed through KNN than to use the H-DIRT method.

Another meaningful matter of fact arising from Table 2 is that the quality of the imputation depends non-negligibly on the experiment. The difference in the error metrics between the various experiments for the same percentage of data clipping can reach order of the 30%. Given this, the analysis per load demand classes is motivated, in order to inquire if there are meaningful trends of the error metrics.

2) ANALYSIS OF THE ACCURACY METRICS PER LOAD DEMAND CLASS

For the top-performing experiment (Experiment 1) detailed in Table 2, metrics per power classes for each percentage of data clipping are depicted in Figures 10, 11, 12, and 13. These figures are based on the pre-processing methods outlined in Section V-C2 and represent outcomes across the varying levels of data clipping examined in the study.

Several observations arise:

- The H-DIRT method has higher MAPE for low power demand classes when the percentage of missing values is low (10%), while the opposite occurs for SKNN.
- In the 10% case, the performance of H-DIRT and SKNN is in average comparable.

TABLE 2. Comparison of metrics for filling missing values in continuous data pruning with results from three experiments with different results scenarios.

Type	% NaN	Metric	Experiment 1			Experiment 2			Experiment 3		
			KNN	SKNN	H-DIRT	KNN	SKNN	H-DIRT	KNN	SKNN	H-DIRT
a_1	10%	RMSE (W)	6687.3	4117.6	4656.1	7356.0	4530.0	5121.7	7420.0	4590.0	5180.0
		MAE (W)	6543.7	3403.4	3680.0	7198.1	3744.0	4048.0	7260.0	3780.0	4100.0
		MAPE (%)	19.4	12.4	12.3	20.3	13.6	13.5	21.5	13.7	13.6
a_2	30%	RMSE (W)	8695.5	4873.1	9134.3	9565.1	5272.1	10026.0	9605.4	5340.0	10030.8
		MAE (W)	8532.3	4814.3	9215.4	9385.5	5296.0	10137.3	9400.6	5301.1	10140.0
		MAPE (%)	26.7	12.1	22.7	29.4	13.3	25.0	29.5	13.4	25.1
a_3	50%	RMSE (W)	9387.6	5852.2	11454.3	10326.0	6437.0	12596.3	10351.1	6450.2	12610.0
		MAE (W)	9211.5	5742.1	11758.8	10133.0	6547.2	12934.0	10155.8	6450.9	12940.0
		MAPE (%)	25.4	14.7	25.4	27.9	16.2	28.0	28.0	16.3	28.1
a_4	70%	RMSE (W)	9011.8	5932.3	14452.7	9913.0	6525.0	15900.0	9890.1	6540.2	15911.0
		MAE (W)	8981.9	6092.4	15580.6	9879.0	6702.0	17240.1	9907.7	6720.5	17150.0
		MAPE (%)	23.4	14.5	36.7	25.8	16.0	37.4	26.0	16.1	39.5

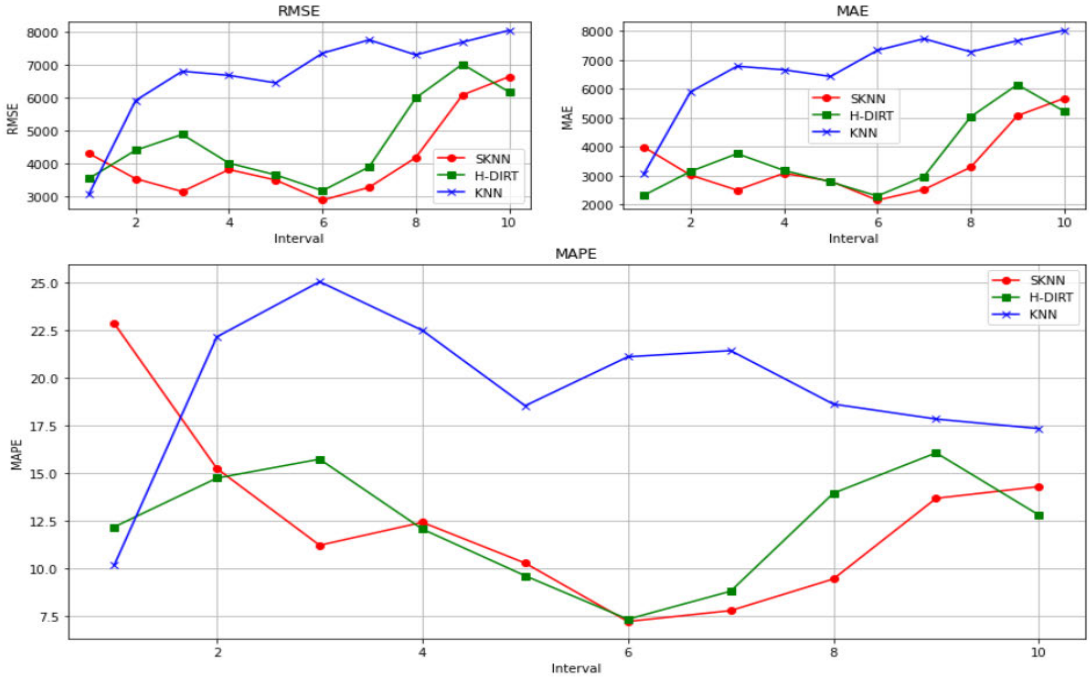

FIGURE 10. Metrics per load demand class for the best experiment (Experiment 1) in the case of 10% data clipping.

TABLE 3. Statistical properties of the measured and imputed data in the case of continuous data clipping: a different random experiment per method is selected.

Type	% NaN	Metric	KNN	Measured	SKNN	Measured	H-DIRT	Measured
a_1	10%	Mean (W)	24962.9	23540.5	29289.8	30944.0	23975.1	23845.3
		Std. Dev. (W)	7341.1	5306.5	7214.3	7748.9	5144.8	5981.3
a_2	30%	Mean (W)	20843.4	26800.0	27506.6	29503.7	28811.2	28783.2
		Std. Dev. (W)	6281.3	8427.5	6988.8	7962.1	10241.0	7974.1
a_3	50%	Mean (W)	22320.8	25631.5	26482.8	29695.6	26033.2	29896.5
		Std. Dev. (W)	8589.5	8893.9	6555.3	8469.1	10918.6	9337.2
a_4	70%	Mean (W)	24902.4	25077.5	26747.2	29694.3	20221.1	28062.1
		Std. Dev. (W)	8403.9	8848.1	6869.1	10928.0	11133.4	8056.6

- The improvement provided by H-DIRT with respect to KNN decreases with the percentage of missing values. Already with 30% of missing values, for several power classes the KNN is better than H-DIRT. This scenario is replicated for the 50% of missing values, while for the 70% case the standard KNN is better than H-DIRT for all the power classes.
- In all the cases except the 10% data clipping, the error metrics for the SKNN case increase with the power demand class but the variability is much lower than for KNN or H-DIRT.
- In all the cases except the 10% data clipping, the SKNN method provides better results than standard KNN and H-DIRT as well.

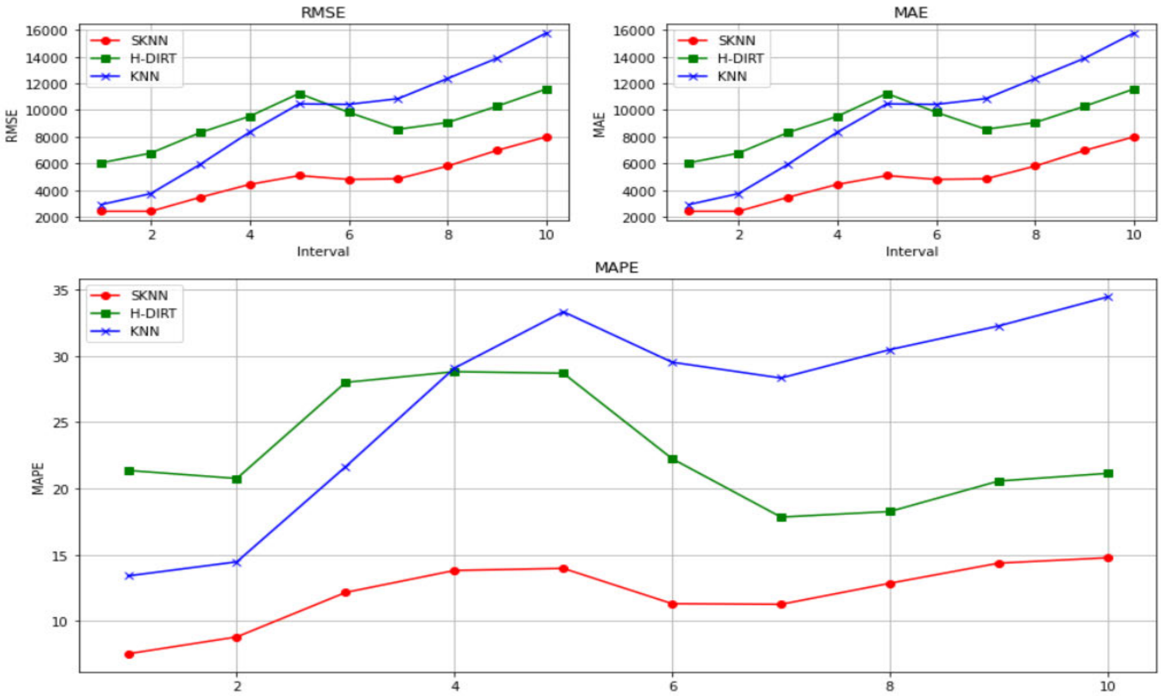

FIGURE 11. Metrics per load demand class for the best experiment (Experiment 1) in the case of 30% data clipping.

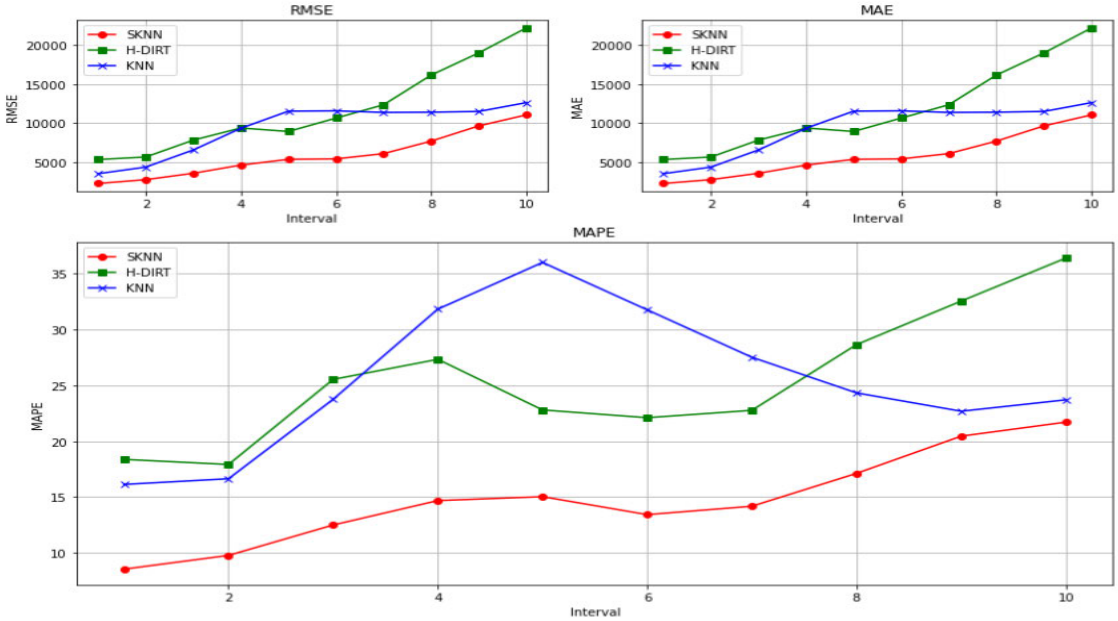

FIGURE 12. Metrics per load demand class for the best experiment (Experiment 1) in the case of 50% data clipping.

- The improvement provided by SKNN with respect to KNN decreases with the increase of the percentage of missing data, but for the considered experiments there is no turning point: the SKNN performs better than KNN for all the power classes for all the experiments.

Given the above results, thus, it can be argued that SKNN provides in general better and stabler results, especially if the percentage of missing measurements is high. In the case of lower percentage of missing values, H-DIRT and SKNN are comparable and the selection of the best method is non-trivial, as there is a non-trivial dependence on the power classes which as a future work will be analyzed more in deep.

3) STATISTICAL ANALYSIS OF THE MEASURED AND IMPUTED DATA

Tables 3 and 4 report the statistical properties of the measured and imputed data, when considering only the clipped data and the whole validation data set respectively. For each

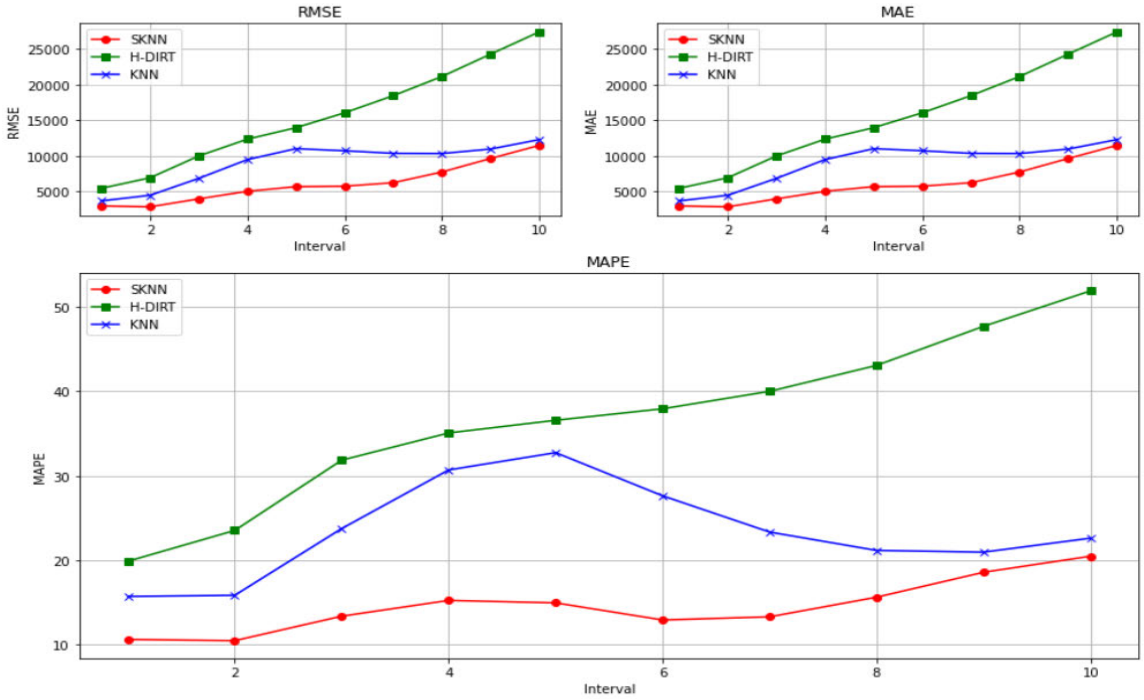

FIGURE 13. Metrics per load demand class for the best experiment (Experiment 1) in the case of 70% data clipping.

TABLE 4. Statistical properties of the whole testing data set with measured and imputed data in the case of continuous data clipping: a different random experiment per method is selected. The measured testing data set has an average of 28096.7 W and a standard deviation of 7906.1 W.

Type	% NaN	Metric	KNN	SKNN	H-DIRT
a_1	10%	Mean (W)	29049.0	28741.1	28480.7
		Std. Dev. (W)	7995.2	7828.1	7867.1
a_2	30%	Mean (W)	26800.0	28172.6	28917.4
		Std. Dev. (W)	8427.5	7652.4	8845.0
a_3	50%	Mean (W)	25631.5	27300.3	26975.1
		Std. Dev. (W)	8893.9	6962.7	9341.6
a_4	70%	Mean (W)	25077.5	26843.7	22975.8
		Std. Dev. (W)	8848.1	7079.0	11793.3

method, the experiment is picked randomly. It should be noted that the same experiments are referred to in both Table 3 and 4, ensuring consistency in the comparison of statistical properties. From such Tables, it is confirmed that SKNN is the method which in general alters less the statistical properties of the time series. Nevertheless, for an acceptably low percentage of clipped data, the H-DIRT method has performance comparable to SKNN. Given this, a further focus on SKNN and H-DIRT for the two extreme cases (10% and 70% of data clipping) is proposed by considering the distributions of measured and imputed data. The results are reported in Figures 14, 15, 16, 17. It arises that SKNN is more brilliant than H-DIRT in predicting the distribution of the imputed data. For example, in the 10% data clipping case (Figure 14), H-DIRT brilliantly captures mean and standard deviation but not the shape of the distribution (which means higher order statistics). In the case of 70% data clipping (Figure 15), H-DIRT even fails to capture brilliantly the average of the time series, which is better predicted by standard KNN. SKNN, instead, manages to reproduce even

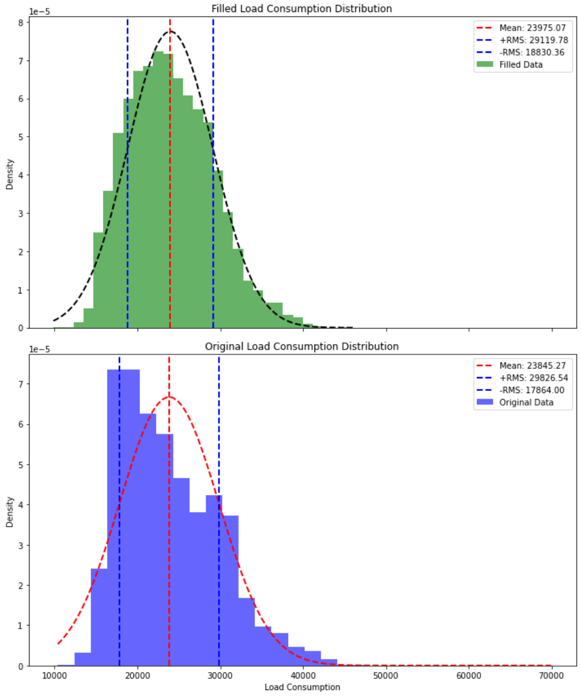

FIGURE 14. The distribution of the continuously clipped data and of the measured data for one experiment in the 10% case for the H-DIRT method.

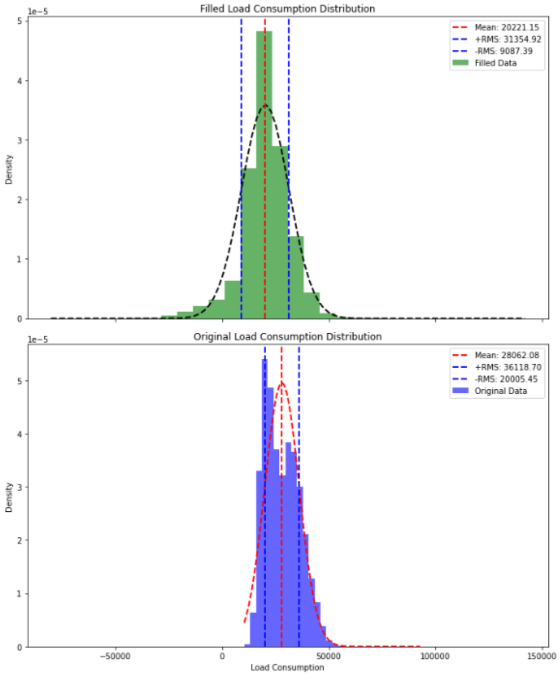

FIGURE 15. The distribution of the continuously clipped data and of the measured data for one experiment in the 70% case for the H-DIRT method.

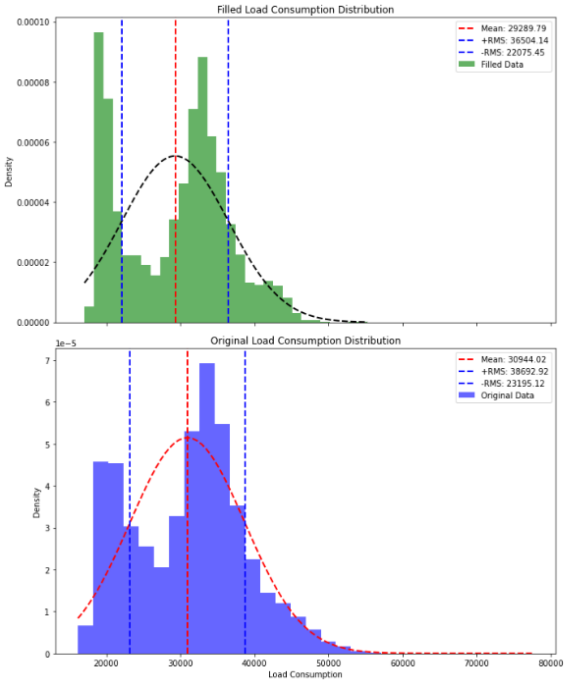

FIGURE 16. The distribution of the continuously clipped data and of the measured data for one experiment in the 10% case for the SKNN method.

non-trivial data distributions, as in the case of the 10% data clipping experiment, where the distribution remarkably deviates from gaussianity and has two peaks.

4) COMPUTATIONAL TIME

In Table 5, the computational time required by each method in the cases of continuous data clipping is reported. It arises that H-DIRT requires 20 to 30 times less time than SKNN. This is a meaningful information, especially for the cases where the two methods provide comparable results in terms of average error metrics.

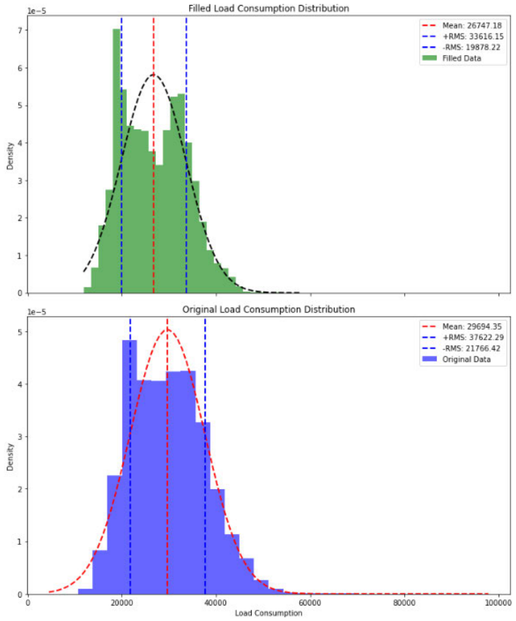

FIGURE 17. The distribution of the continuously clipped data and of the measured data for one experiment in the 70% case for the SKNN method.

TABLE 5. Computation times (in seconds) for different methods at various levels of data pruning.

Method	10%	30%	50%	70%
HDIRT	11	31	52	127
KNN	20	51	121	375
SKNN	368	1087	1743	2224

VII. CONCLUSION

The present work deals with the formulation of innovative methods for data imputation, which can especially be useful in the context of time series analysis.

The proposed methods are aimed at conjugating simplicity (and, thus, interpretability) with accuracy. They substantially leverage state of the art methods (like KNN or Linear Regression data imputation) and improve them by judiciously incorporating season-related information.

The main building block of the proposed H-DIRT method is a multivariate linear regression, which works between the power demand measurement and an appropriately constructed vector of historical power demands. The imputation then proceeds by employing such model if the historical measurements are available. In case the historical measurements are partially available, a data averaging is employed. If the requested historical measurements are totally unavailable, a linear interpolation is carried out between the first available measurements before and after the missing value.

The SKNN method proposed in this work consists of the baseline KNN employing the K nearest neighbors measurement in the time domain, upon feature enrichment, by adding yearly, seasonal, monthly, weekly and daily information.

A qualifying aspect of the work is that the proposed methods have been tested on a real-world data set of several years, from an electric power station situated on the engineering campus of the University of Brescia, Italy.

The results, collected in Section VI, in general indicate that both methods are remarkably more accurate with respect to the state of the art KNN imputation method applied in the time domain. The takeaway message from this work is that it is possible to perform data imputation, even in challenging contexts as high shares of continuous missing data, by employing methods which building blocks are simple state of the art ones, but enriched with seasonal information, which for SKNN and H-DIRT is included in two different ways. Which between H-DIRT or SKNN works better depends on the context, but in general it can be concluded that SKNN provides stabler and more accurate results. Nevertheless, especially in the case of continuous data clipping with a not too high percentage of missing data, the performance of H-DIRT was comparable to that of SKNN and, depending on the application, it might be more favorable, given that it requires from 20 to 30 less times computational time.

There are possible further directions of this work. One of the most interesting regards how to combine the separate predictions from H-DIRT and SKNN, as for example through decision trees or stacking. Further effort by the authors is at present being devoted to the analysis of how the data imputation impacts on the short-term load forecasting and on the statistical analysis of the observed load trends. Furthermore, it should be noticed that the proposed methods are in line of principle effective also for dealing with bad data imputation [42], but this entails further elaboration in order to determine how bad data are defined, which was beyond the scope of the present paper.

REFERENCES

[1] M. González-Torres, L. Pérez-Lombard, J. F. Coronel, I. R. Maestre, and D. Yan, "A review on buildings energy information: Trends, end-uses, fuels and drivers," *Energy Rep.*, vol. 8, pp. 626–637, Nov. 2022.

[2] L. Zhang, J. Wen, Y. Li, J. Chen, Y. Ye, Y. Fu, and W. Livingood, "A review of machine learning in building load prediction," *Appl. Energy*, vol. 285, Mar. 2021, Art. no. 116452.

[3] M. Pazhoohesh, Z. Pourmirza, and S. Walker, "A comparison of methods for missing data treatment in building sensor data," in *Proc. IEEE 7th Int. Conf. Smart Energy Grid Eng. (SEGE)*, Aug. 2019, pp. 255–259.

[4] P. Golovinski, D. Vasenin, N. Savvin, S. Rinaldi, and M. Pasetti, "Electricity consumption forecast of clusters of buildings based on recurrent neural networks," in *Proc. 4th Int. Conf. Control Syst., Math. Model., Autom. Energy Efficiency (SUMMA)*, Nov. 2022, pp. 375–380.

[5] B. Cho, T. Dayrit, Y. Gao, Z. Wang, T. Hong, A. Sim, and K. Wu, "Effective missing value imputation methods for building monitoring data," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2020, pp. 2866–2875.

[6] D. Mariano-Hernández, L. Hernández-Callejo, A. Zorita-Lamadrid, O. Duque-Pérez, and F. Santos García, "A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis," *J. Building Eng.*, vol. 33, Jan. 2021, Art. no. 101692.

[7] P. B. Weerakody, K. W. Wong, G. Wang, and W. Ela, "A review of irregular time series data handling with gated recurrent neural networks," *Neurocomputing*, vol. 441, pp. 161–178, Jun. 2021.

[8] Y. Zhou, L. Ma, W. Ni, and C. Yu, "Data enrichment as a method of data preprocessing to enhance short-term wind power forecasting," *Energies*, vol. 16, no. 5, p. 2094, Feb. 2023.

[9] Z. Zhang, "Missing data imputation: Focusing on single imputation," *Ann. Transl. Med.*, vol. 4, no. 1, p. 9, 2016.

[10] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of performance of data imputation methods for numeric dataset," *Appl. Artif. Intell.*, vol. 33, no. 10, pp. 913–933, Aug. 2019.

[11] K. Seu, M.-S. Kang, and H. Lee, "An intelligent missing data imputation techniques: A review," *JOIV : Int. J. Informat. Visualizat.*, vol. 6, nos. 1–2, pp. 278–283, May 2022.

[12] J. Josse, J. Pagès, and F. Husson, "Multiple imputation in principal component analysis," *Adv. Data Anal. Classification*, vol. 5, no. 3, pp. 231–246, Oct. 2011.

[13] R. Zhai and R. Gutman, "A Bayesian singular value decomposition procedure for missing data imputation," *J. Comput. Graph. Statist.*, vol. 32, no. 2, pp. 470–482, Apr. 2023.

[14] D. Jeong, C. Park, and Y. M. Ko, "Missing data imputation using mixture factor analysis for building electric load data," *Appl. Energy*, vol. 304, Dec. 2021, Art. no. 117655.

[15] M. N. Noor, A. S. Yahaya, N. A. Ramli, and A. M. M. Al Bakri, "Filling missing data using interpolation methods: Study on the effect of fitting distribution," *Key Eng. Mater.*, vols. 594–595, pp. 889–895, Dec. 2013.

[16] T. Liu, H. Wei, and K. Zhang, "Wind power prediction with missing data using Gaussian process regression and multiple imputation," *Appl. Soft Comput.*, vol. 71, pp. 905–916, Oct. 2018.

[17] F. Bengtsson and K. Lindblad, "Methods for handling missing values: A simulation study comparing imputation methods for missing values on a Poisson distributed explanatory variable," Univ. Uppsal, Uppsala, Sweden, Tech. Rep., 2020.

[18] W.-C. Lin and C.-F. Tsai, "Missing value imputation: A review and analysis of the literature (2006–2017)," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1487–1509, Feb. 2020.

[19] C. E. Borges, O. Kamara-Esteban, T. Castillo-Calzadilla, C. M. Andonegui, and A. Alonso-Vicario, "Enhancing the missing data imputation of primary substation load demand records," *Sustain. Energy, Grids Netw.*, vol. 23, Sep. 2020, Art. no. 100369.

[20] A. Chong, K. P. Lam, W. Xu, O. T. Karaguzel, and Y. Mo, "Imputation of missing values in building sensor data," *ASHRAE IBPSA-USA SimBuild*, vol. 6, p. 407, Aug. 2016.

[21] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, Mar. 2010.

[22] M. Weber, M. Turowski, H. K. Çakmak, R. Mikut, U. Kühnapfel, and V. Hagenmeyer, "Data-driven copy-paste imputation for energy time series," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5409–5419, Nov. 2021.

[23] D. Magare, S. Labde, M. Gofane, and V. Vyawahare, "Imputation of missing data in time series by different computation methods in various data set applications," in *Proc. ITM Web Conf.*, vol. 32, 2020, p. 03010.

[24] S. Alonso, A. Morán, D. Pérez, M. A. Prada, J. J. Fuertes, and M. Domínguez, "Gap imputation in related multivariate time series through recurrent neural network-based denoising autoencoder1," *Integr. Comput.-Aided Eng.*, vol. 31, no. 2, pp. 157–172, 2024.

[25] T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting," *Appl. Sci.*, vol. 9, no. 1, p. 204, Jan. 2019.

[26] A. Wang, J. Yang, and N. An, "Regularized sparse modelling for microarray missing value estimation," *IEEE Access*, vol. 9, pp. 16899–16913, 2021.

[27] K. Psychogyios, L. Ilias, C. Ntanos, and D. Askounis, "Missing value imputation methods for electronic health records," *IEEE Access*, vol. 11, pp. 21562–21574, 2023.

[28] A. Lucbert, J. van der Niet, A. Corson, M. Weij, R. I. van der Elst, J. M. M. de Juan, and T. B. S. Rahola, "Time series building energy systems data imputation," in *Proc. CLIMA Conf.*, 2022, pp. 1–8.

[29] R. Shahbazian and S. Greco, "Generative adversarial networks assist missing data imputation: A comprehensive survey and evaluation," *IEEE Access*, vol. 11, pp. 88908–88928, 2023.

[30] W. Khan, N. Zaki, A. Ahmad, M. M. Masud, L. Ali, N. Ali, and L. A. Ahmed, "Mixed data imputation using generative adversarial networks," *IEEE Access*, vol. 10, pp. 124475–124490, 2022.

[31] S.-Y. Lee, T. P. Connerton, Y.-W. Lee, D. Kim, D. Kim, and J.-H. Kim, "Semi-GAN: An improved GAN-based missing data imputation method for the semiconductor industry," *IEEE Access*, vol. 10, pp. 72328–72338, 2022.

[32] A. Ijadi Maghsoodi, A. E. Torkayesh, L. C. Wood, E. Herrera-Viedma, and K. Govindan, "A machine learning driven multiple criteria decision analysis using LS-SVM feature elimination: Sustainability performance assessment with incomplete data," *Eng. Appl. Artif. Intell.*, vol. 119, Mar. 2023, Art. no. 105785.

[33] M.-C. Wang, C.-F. Tsai, and W.-C. Lin, "Towards missing electric power data imputation for energy management systems," *Exp. Syst. Appl.*, vol. 174, Jul. 2021, Art. no. 114743.

[34] A. Lotfipoor, S. Patidar, and D. P. Jenkins, "Transformer network for data imputation in electricity demand data," *Energy Buildings*, vol. 300, Dec. 2023, Art. no. 113675.

[35] X. Liu, X. Lai, and L. Zhang, "A hierarchical missing value imputation method by correlation-based K-nearest neighbors," in *Proc. Intell. Syst. Conf. (IntelliSys)*, vol. 1. London, U.K.: Springer, 2020, pp. 486–496.

[36] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *J. Syst. Softw.*, vol. 85, no. 11, pp. 2541–2552, Nov. 2012.

[37] N. Karmitsa, S. Taheri, A. Bagirov, and P. Mäkinen, "Missing value imputation via clusterwise linear regression," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1889–1901, Apr. 2022.

[38] A. Flammini, M. Pasetti, S. Rinaldi, P. Bellagente, A. C. Ciribini, L. C. Tagliabue, L. E. Zavanella, S. Zanoni, G. Oggioni, and G. Pedrazzi, "A living lab and testing infrastructure for the development of innovative smart energy solutions: The eLUX laboratory of the University of Brescia," in *Proc. AEIT Int. Annu. Conf.*, Oct. 2018, pp. 1–6.

[39] M. Pasetti, "Assessing the effectiveness of the energy storage rule-based control in reducing the power flow uncertainties caused by distributed photovoltaic systems," *Energies*, vol. 14, no. 8, p. 2312, Apr. 2021.

[40] H. Daneshi and A. Daneshi, "Real time load forecast in power system," in *Proc. 3rd Int. Conf. Electric Utility Deregulation Restructuring Power Technol.*, Apr. 2008, pp. 689–695.

[41] C. Guan, P. B. Luh, L. D. Michel, Y. Wang, and P. B. Friedland, "Very short-term load forecasting: Wavelet neural networks with data pre-filtering," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 30–41, Feb. 2013.

[42] J. Peppanen, X. Zhang, S. Grijalva, and M. J. Reno, "Handling bad or missing smart meter data through advanced data imputation," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Sep. 2016, pp. 1–5.

DMITRII VASENIN (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in civil engineering from Voronezh State Technical University, Russia. He is currently pursuing the Ph.D. degree in technology for health with the Department of Information Engineering, University of Brescia. His current research interests include time series forecasting, renewable energy, smart grids, and data analysis.

MARCO PASETTI (Member, IEEE) received the M.Sc. degree in industrial engineering and the Ph.D. degree in mechanical engineering from the University of Brescia, Brescia, Italy, in 2008 and 2013, respectively. He is currently an Assistant Professor in electrical energy systems with the Department of Information Engineering, University of Brescia. His current research interests include energy systems, distributed generation, renewable energy sources, photovoltaics, energy storage, demand-side management, electric vehicles charging systems, energy management systems, supervisory control and data acquisition, and smart grids.

DAVIDE ASTOLFI (Member, IEEE) received the B.S., M.S., and Ph.D. degree in physics the Ph.D. degree in industrial and information engineering from the University of Perugia, Italy. He is currently an Assistant Professor in electrical systems for energy with the University of Brescia. His research interests include wind turbine technology, data analysis and artificial intelligence applications for renewable energy, supervisory control and data acquisition, load and renewable power forecasting, and electric vehicles. He is a Subject Editor of the *IET Renewable Power Generation* journal and an Associate Editor of *Smart Grid and Sustainable Energy* journal.

NIKITA SAVVIN received the B.S. degree in thermal power engineering and the M.S. degree in artificial intelligence technologies from Voronezh State Technical University (VSTU), Voronezh, Russia, where he is currently pursuing the Ph.D. degree with the Department of Automated Computing Systems, Program of System Analysis, Information Management and Processing. He is also a Programmer with the HSE Institute for Statistical Research and Knowledge Economics, Moscow, Russia. His current research interests include time series forecasting, neural language processing, and data analysis.

STEFANO RINALDI (Senior Member, IEEE) received the degree (Hons.) in electronic engineering and the Ph.D. degree in electronic instrumentation from the University of Brescia, Brescia, Italy, in 2006 and 2010, respectively. He is currently an Associate Professor with the Department of Information Engineering, University of Brescia. His research interests include industrial real-time ethernet networks, the Internet of Things, time synchronization, smart grids, renewable energy sources, electric vehicles, and cognitive building.

ALBERTO BERIZZI (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Politecnico di Milano, in 1994. He is currently a Full Professor in electric power systems with the Politecnico di Milano. He has authored more than 200 scientific articles and has been responsible for many research projects, funded both by institutions and by the industry. His research interests include power system analysis, security, optimization, and control.

Open Access funding provided by 'Università degli Studi di Brescia' within the CRUI CARE Agreement