POLITECNICO DI TORINO
Repository ISTITUZIONALE

A data-driven approach to predict hourly load profiles from time-of-use electricity bills

Original
A data-driven approach to predict hourly load profiles from time-of-use electricity bills / Lazzeroni, Paolo; Lorenti, Gianmarco; Repetto, Maurizio. - In: IEEE ACCESS. - ISSN 2169-3536. - 11:(2023). [10.1109/ACCESS.2023.3286020]

Availability:
This version is available at: 11583/2979386 since: 2023-06-15T09:05:53Z

Publisher:
IEEE

Published
DOI:10.1109/ACCESS.2023.3286020

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

30 June 2024

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.0122113

A data-driven approach to predict hourly load profiles from time-of-use electricity bills

PAOLO LAZZERONI[1], GIANMARCO LORENTI[1], and MAURIZIO REPETTO[1]
[1]Politecnico di Torino, Dipartimento di Energia "Galileo Ferraris", Corso Duca degli Abruzzi, 24, Torino 10129, Italia (e-mail: name.surname@polito.it)

Corresponding author: Gianmarco Lorenti (e-mail: gianmarco.lorenti@polito.it).

This work was supported by Project 'NEXT GENERATION WE', funded by Fondazione Compagnia di San Paolo, Torino, Italy.

ABSTRACT The design of renewable-based and collective energy systems requires consumption data with fine temporal and spatial resolution. Despite the increasing deployment of smart meters, obtaining such data directly from measurements can still be challenging, particularly when attempting to gather historical data over a reasonable period for many end users. As a result, there is a need for models to simulate or predict these consumption data (e.g., hourly load profiles). Typically, these models rely on numerous specific and detailed observations, such as load type, household size for residential customers, or operating hours for commercial ones. However, gathering this level of detail becomes increasingly difficult as the number and diversity of end users increase. Therefore, this paper proposes a data-driven approach to predict hourly load profiles of heterogeneous end users using only their monthly time-of-use electricity bills as inputs. We create a training set using a limited number of hourly measurements from diverse categories of end users and, differently from other approaches aimed at classifying the end users, we develop a regression model to map monthly electricity bills to typical load profiles. Experimental results using one year of data from various end user categories demonstrate an average normalized mean absolute error of approximately 26% for instantaneous consumption and less than 4% for time-of-use values. Comparative analysis with standard load profiles and a two-step data-driven approach based on classification reveals that our proposed method outperforms the others in terms of prediction accuracy and statistical metrics.

INDEX TERMS
Energy consumption, electricity demand, load modeling, data-driven modeling, nearest neighbor methods.

I. INTRODUCTION

THE decarbonization of electricity will play a relevant role in the future, worldwide energy transition to transform the electricity grid into less carbon-intense or, potentially, carbon-free systems [1]. This transition will require more Renewable Energy Sources (RESs) to be integrated into the grid at both utility-scale and the small and local scale on a more dispersed perspective, as distributed generation [2]. The former, as large-size installations, usually participate in the wholesale markets, and hence their sizing does not dependent on local energy consumption. The latter solutions, instead, sustain local demand by increasing self-sufficiency and reducing supply costs in many fields of application, from the industrial [3] to the residential sectors [4].

Hence, RES-based distributed generation must be properly sized according to the energy demand, to make the installations affordable, profitable, and sustainable. Practically, the match between production and consumption at fine temporal resolution (at least hourly [5]) must be thoroughly assessed, e.g., through optimization [6], to avoid potential over-sizing or underestimation, regardless of the field of application.

For this reason, weather data are needed to estimate RES production through proper modeling of the active assets [7], while load (i.e., consumption) profiles of the local demand need to be identified [8]. This last task can become complex when the number and the typology of end users increase significantly, due to limited access to data measured by smart meters (SMs) and the conse-

quent increase in the number of unobserved end users (i.e., whose consumption data cannot be collected at fine temporal resolution). In fact, collecting hourly (or sub-hourly) consumption data from SM can be challenging due to various factors, including privacy concerns and technical limitations. Privacy regulations often restrict the collection of granular consumption data, as they can reveal sensitive information about individuals' activities and habits. Moreover, the roll out of new SMs to allow measurements at fine temporal granularity for a large number of customers can be logistically demanding and time-consuming [9], [10]. Consequently, hourly consumption profiles are not always available. Energy bills, instead, provide aggregated consumption information and are typically already collected as part of standard billing procedures, making them readily available directly to the end users or suppliers. Despite being at a coarse temporal resolution, energy bills data provide valuable information for the analysis and prediction of load profiles, e.g., through data-driven modeling.

Unlike standard load profiles (SLPs) [11] or synthetically generated ones [12], data-driven approaches use measurements to build a model that relates attributes of the end users to their load profiles. This knowledge can then be extrapolated to predict load profiles of unobserved customers. Previous studies [13], [14] have utilized different attributes as input data. However, these approaches often require intensive and intrusive data collection, whereas the energy bill offers a potential solution by enabling models that can handle limited data [15].

Therefore, in this study, we propose a data-driven approach to predict the electricity load profiles of unobserved end users using only their electricity bills. Specifically, we leverage time-of-use (ToU) tariffs [16], [17], which divide the monthly bill based on varying consumption patterns throughout the day and week. Then, by applying a data-driven approach, our goal is to invert the relationship between the ToU consumption and the load profile that generated it. While it may be argued that different load profiles can result in the same energy bill, due to its integral nature, practical observations suggest a limited number of distinct consumption patterns [18]. We assume that these identifiable patterns are associated with different energy bills. Unlike classification-based approaches [13]–[15], we utilize a regression model, specifically k-nearest neighbors (k-NN), to map energy bills to hourly load profiles.

Notwithstanding its limitations, the proposed approach is easily exploitable in contexts where managing a portfolio of end users is required, including unobservable ones. In fact, once the model is trained using a limited amount of measurements, it only utilizes aggregated data from monthly energy bills, which can be collected even for large numbers of heterogeneous end users. The fields of application of the approach extend beyond RES-based local energy systems to encompass the design and evaluation of Demand Response programs, identification of energy-saving opportunities, and creation of dedicated offers from energy suppliers.

The remaining sections are structured as follows. In Section II, we review the literature on load modeling, focusing on data-driven methods and presenting the related works and contributions of this study. Section III analyzes the load prediction from energy bills as an inverse problem. Section IV describes the implementation methods for the proposed approach and benchmark approaches from the literature. Section V discusses the data set used for testing the proposed approach. Section VI presents and compares the results obtained with benchmark approaches, discussing the strengths and limitations of the proposed approach. Finally, Section VII provides the conclusions of this work.

II. LITERATURE REVIEW AND CONTRIBUTION

A. LOAD MODELING

This work pertains to the field of energy use/load modeling [19], specifically focusing on the prediction of fine-resolution electricity consumption data for heterogeneous end users (residential, commercial, public offices).

Traditionally, SLPs obtained from measurement campaigns have been widely utilized. For example, the H0 SLP is employed by energy companies in Germany and Austria for residential customers [11], [12]. Similarly, non-residential customers are categorized and assigned an SLP based on their assumed energy usage and/or energy intensity [20], [21]. However, SLPs have faced criticism due to their reliance on outdated data, ignoring recent changes in consumption patterns, and overlooking variations within the same category [12], [20], [22]. Moreover, SLPs struggle to accurately capture consumption levels and dynamics at fine temporal and spatial resolutions [11], [23]. Nevertheless, SLPs continue to be widely used in industry and government bodies, such as the Italian Manager of Energy Services (GSE), which adopted SLPs for regulating end users under collective self-consumption and energy communities to address the absence of SM data [24].

Conversely, research has focused on modeling energy usage to generate synthetic load profiles. Various models, classified as bottom-up and top-down (sometimes hybrid), have been proposed [25]. Extensive reviews of these models have been conducted by Grandjean et al. [26] and Proedrou [12]. In essence, bottom-up models can produce accurate and highly resolved load profiles but require extensive and detailed input data, while top-down models can work with fewer, large-scale data, but they are generally more suitable for aggregate-level load

profile modeling [12], [25][1].

Recently, data-driven models have flourished in the literature thanks to the introduction of Advanced Metering Infrastructure, of which SMs are a key component [28]–[30]. These models typically treat hourly or sub-hourly energy consumption data from SMs as time series and employ data mining and machine learning techniques to extract knowledge from them [31]. While these models find applications in various areas such as fault and anomaly/bad data detection, and load forecasting [32], our focus is on the topic of load management [29], which encompasses building energy benchmarking and customer segmentation.

A significant portion of the literature is focused on the latter field, which aims to group energy end users based on similar characteristics. One approach is to cluster buildings with similar energy usage patterns to identify energy-saving opportunities [33]–[35]. Another approach involves identifying groups of customers with similar consumption patterns, enabling the provision of tailored offers and the implementation of Demand Response programs to enhance energy system operation [13]–[15], [22]. Studies by Rasanen et al. [20], Mutanen et al. [36], and, more recently, Parks et al. [18] and Zhan et al. [37] demonstrate how data-driven methods analyzing SM data yield more accurate results compared to traditional approaches based on activity typology (e.g., residential, commercial, government) and energy use intensity.

Clustering SM consumption data is a common characteristic of these data-driven methods, which involves unsupervised learning [38] to identify groups with similar consumption patterns. Each group, or cluster, is then associated with a representative load profile (i.e., the cluster center). Chicco [28], Zhou et al. [39] and Rajabi et al. [21] provide comprehensive reviews of load profile clustering, covering the methodologies, steps, application to customer classification, and performance evaluation metrics. It is important to note that only a few cases include a post-clustering phase, which entails supervised learning [38] to develop models such as classifier systems or regressions for extrapolating the acquired knowledge to new (or unobservable) customers.

B. RELATED DATA-DRIVEN APPROACHES

The post-clustering phase serves two opposing aims: inferring the characteristics of buildings/end users from their load profiles (used as predictive attributes) [40], [41]; predicting the class (and, consequently, load profile) of new or unobservable customers based on a series of attributes. The latter problem, on which we focus, has received little attention in the literature according to previous studies [13], [15] and to our own review.

[1]It is worth noting that Duque et al. [27] recently developed a probabilistic model using smart meter measurements to generate synthetic load profiles for individual households, conditioned on specific total yearly consumption values.

Typically, the classification of new customers and the assignment of representative profiles rely on specific attributes of the end users within each cluster. For instance, Viegas et al. [14] utilized survey data, such as the age and income of family members, and the number of appliances, along with a limited amount of SM measurements (0 to 10 weeks). In the case of unobserved end users, the classification solely relies on the available survey data. Vercamer et al. [13], instead, classified non-residential consumers based on internal company data (e.g., commercial code, number of employees), open data related to the municipality, and cartographic information (e.g., building size). Piscitelli et al. [15], on the other hand, classified commercial and industrial end users based on easily collectible data such as monthly consumption from energy bills and information on opening/closing and lunchtime hours. In contrast to the aforementioned methods, Granell et al. [42] employed a regression approach using k-NN to predict the hourly load profiles of new supermarkets in different market areas, considering the floor area.

Overall, previous models have been developed based on category-specific end user attributes or through extensive data collection procedures. However, these approaches may not be suitable when dealing with a large number of end users from diverse categories, as they can become cumbersome and inefficient.

C. CONTRIBUTION

This paper introduces different key contributions to the field of load modeling and, in particular, end user load profile prediction. By building upon the existing literature, we address the following aspects.

Utilization of ToU energy bills

The analyzed state-of-the-art approaches often rely on end-user attributes that are category-specific or require intensive data collection, which can be cumbersome and impractical. Conversely, our approach utilizes only the ToU energy bill for load profile prediction. While Piscitelli et al. [15] also used energy bills, their approach incorporated additional data, specific to commercial and industrial customers. Furthermore, they focused solely on weekdays, while we consider a more comprehensive set of typical days (also including Saturdays, and holidays, according to the ToU tariffs described in Section III).

Regression-based prediction

Unlike classification-based approaches employed in previous works, we adopt a regression-based methodology to predict hourly load profiles. We frame the problem as an inverse one since the input (energy bill) and output (consumption profile) are linked by a direct analytical relationship. SLPs, commonly used in the industry and by government bodies [24], map energy bills to load

profiles independently of the bill's characteristics. In other words, the same profile is indifferently adopted for all end users (within a predefined category), while re-scaling is used to preserve the total consumption (see Fig. 1a). Alternatively, clustering-classification approaches map different energy bills to distinct load profiles (i.e., the cluster centers) which can be seen as a "discrete" mapping (see Fig. 1b). In contrast, our approach leverages regression using k-NN, enabling a more continuous mapping between energy bills and load profiles (see Fig. 1c). In this way, similar ToU energy bills are mapped into similar load profiles, reflecting the supposed inverse relationship between the two quantities, which is described more in detail in Section III.

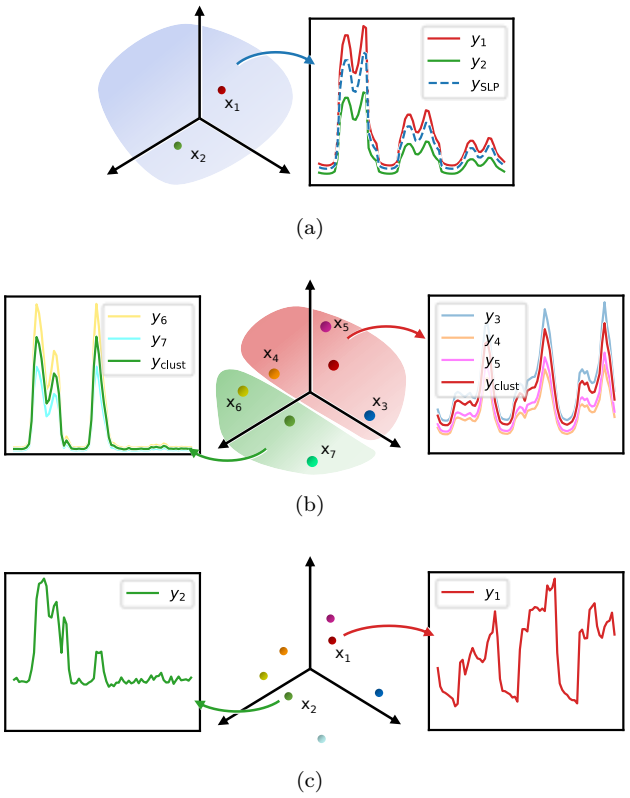

FIGURE 1. **Visual representation exemplifying different ways of mapping the energy bills (represented as points in the x-space) to their corresponding typical load profile (i.e., their image in the y-space), which is represented as a line (see Section III-A for the definitions of these elements). These three methods of mapping are compared in the paper: in (a) an SLP is used, and consequently the values of an energy bill can only change the total consumption ("magnitude") of the typical load profile while the shape corresponds to the SLP; differently, in (b), the x-space is divided into clusters, and hence depending on the values of the energy bill (i.e., the position of a point in the x-space), it is mapped by with a different typical load profile whose shape corresponds to the cluster center; finally, in (c), each point is considered individually and mapped in a more continuous way, in a regression-like approach, by means of k-NN.**

Testing diverse end-user categories

We leverage the fact that our proposed approach provides a more versatile and scalable solution that can be applied to a wide range of end-user categories. Therefore, while previous studies focus on a single category, we test the proposed approach on different customer categories, involving residential, commercial, and public offices. We also provide a comparison with two benchmarks, respectively based on SLPs and the two steps clustering-classification approach identified in the literature.

In summary, these contributions advance the field of load modeling and prediction, in particular by providing a streamlined and data-driven approach that utilizes time-of-use energy bills for regression-based load profile prediction.

III. INVERSE PROBLEM DESCRIPTION

We start by providing an overview of the ToU tariff scheme implemented in the Italian regulation [43] since we use a set of SM measurements of Italian end users for testing our approach. In particular, three different tariffs are defined for electricity (F1, F2, F3), respectively, for on-peak, mid-peak, and off-peak hours, which are arranged as shown in Fig. 2. According to these ToU tariffs, three types of days can be identified, each characterized by a different subdivision of the hours into ToU tariffs: work days, from Monday to Friday; Saturdays; and Sundays/holidays (just 'holidays', in the followings).

A. DEFINITIONS

Load profile

We refer to a load profile as a time series of the energy consumption over a sequence of uniformly-spaced time steps, which in this paper have an hourly resolution. We use an average, uniform power demand associated with the energy consumption in each time step while using the term 'energy' for other quantities with rougher granularity (e.g., daily or monthly).

Average load profile

An average load profile is the result of a time step-by-time step average between load profiles that have the same length and that share some feature(s). In this paper, we evaluate average load profiles on a monthly basis, for each day type (work days, Saturdays, holidays).

Typical load profile

A typical load profile provides a condensed representation of the three average load profiles in a month, obtained by putting the latter in a sequence (see Fig. 3).

Energy bill

An energy bill is a set of records of the energy consumption (rather than the expenditure) in the three ToU tariffs in one month. The components of the energy bill can be evaluated from a typical load profile, considering the ToU structure depicted in Fig. 2.

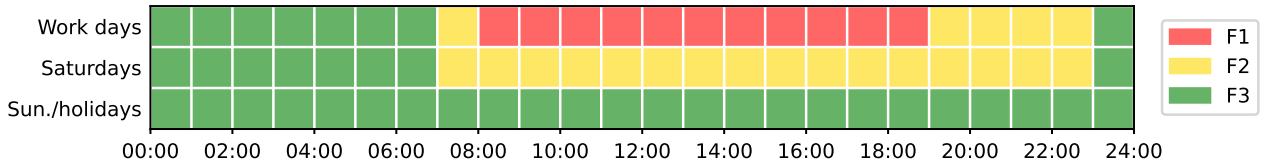

FIGURE 2. ToU tariffs structure adopted in Italy. The central hours of work days are on-peak (tariff F1, in red); early morning and evening of work days (Monday-Friday) and day hours of Saturdays are mid-peak (tariff F2, in yellow); night, Sundays, and major holidays are off-peak hours (tariff F3, in green).

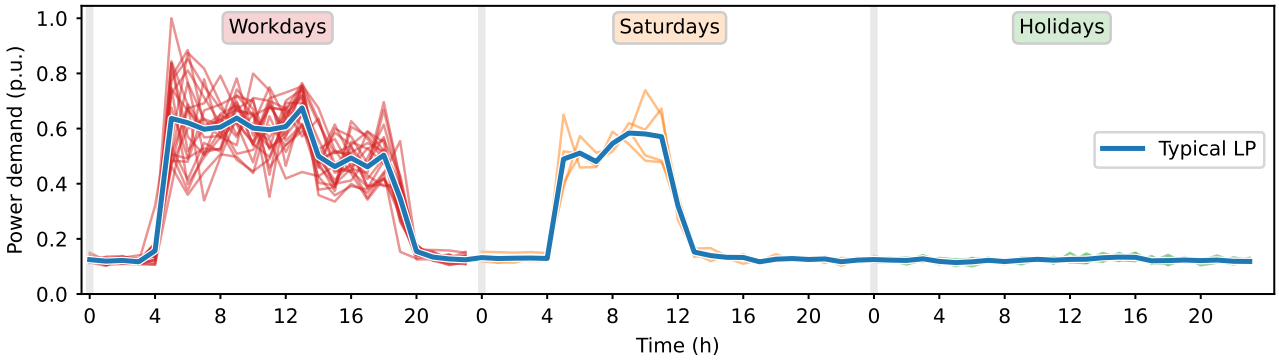

FIGURE 3. Visual example of the averaging process and creation of a typical load profile in one month. A month-long load profile is first chunked into day-long sequences, which are arranged by day type (in red are the load profiles of all work days, in yellow those of all Saturdays, and in green holidays); then, an average load profile is evaluated for each day type; these profiles are finally put in a sequence to obtain the typical load profile (the blue, continuous line). [The data are from the data set described in Section V].

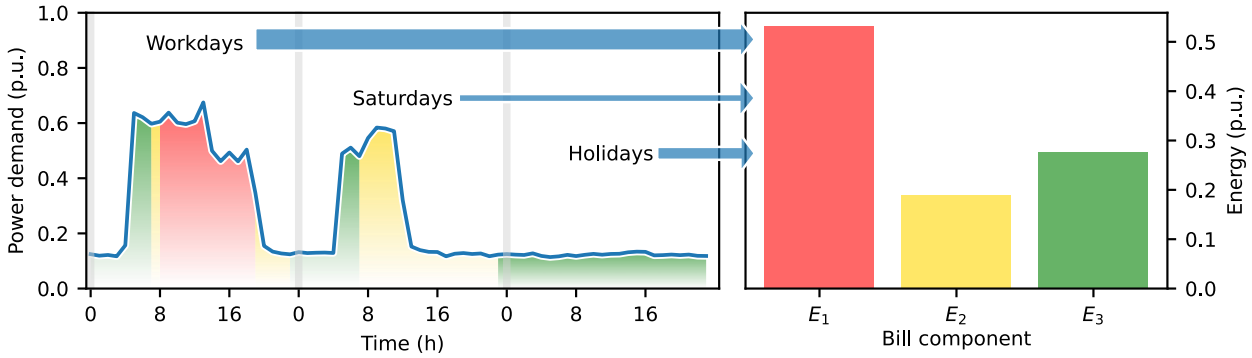

FIGURE 4. Visual example of the bill calculation from a typical load profile in one month. The blue line shows the typical load profile. The subdivision of the time steps into ToU tariffs is also shown (F1, in red, F2, in yellow, and F3, in green). The colored bars are the three components of the bill, i.e., the consumption in each ToU tariff. The width of the arrow is related to the number of days of each day type in the month.

Spaces of the energy bills and typical load profiles

An energy bill can be imagined as a point in a space with three dimensions, i.e., one for each ToU tariff: we call this the 'x-space'. Similarly, we define a space of the typical load profiles, the 'y-space', that has one dimension for each time step in the profile.

B. DIRECT AND INVERSE MAPPING

Given the definitions provided above, the calculation of the ToU energy bill from a typical load profile provides the direct analytical relationship between the two quantities. Fig. 4 provides a qualitative example of this relationship, while the mathematical details are reported in Section IV. On the other hand, the analysis of measured data shows that it is possible to invert this relationship. For instance, in Fig. 5, three different pairs of energy bills are shown in the x-space, with the related typical load profiles in the y-space. These have been randomly chosen among the data available in the testing data set (described in Section V) in order to be pair-wise close in the bills space although from different end users. Despite some differences in the instantaneous consumption within the same ToU tariff time span, the profiles that share similar energy bills also share many similarities in the consumption patterns, e.g., the hours of peak and base load, the number of spikes in the consumption, and so on. In the following section, the methods to leverage this inverse relationship to build a model for the prediction of typical load profiles of unobservable end users are described.

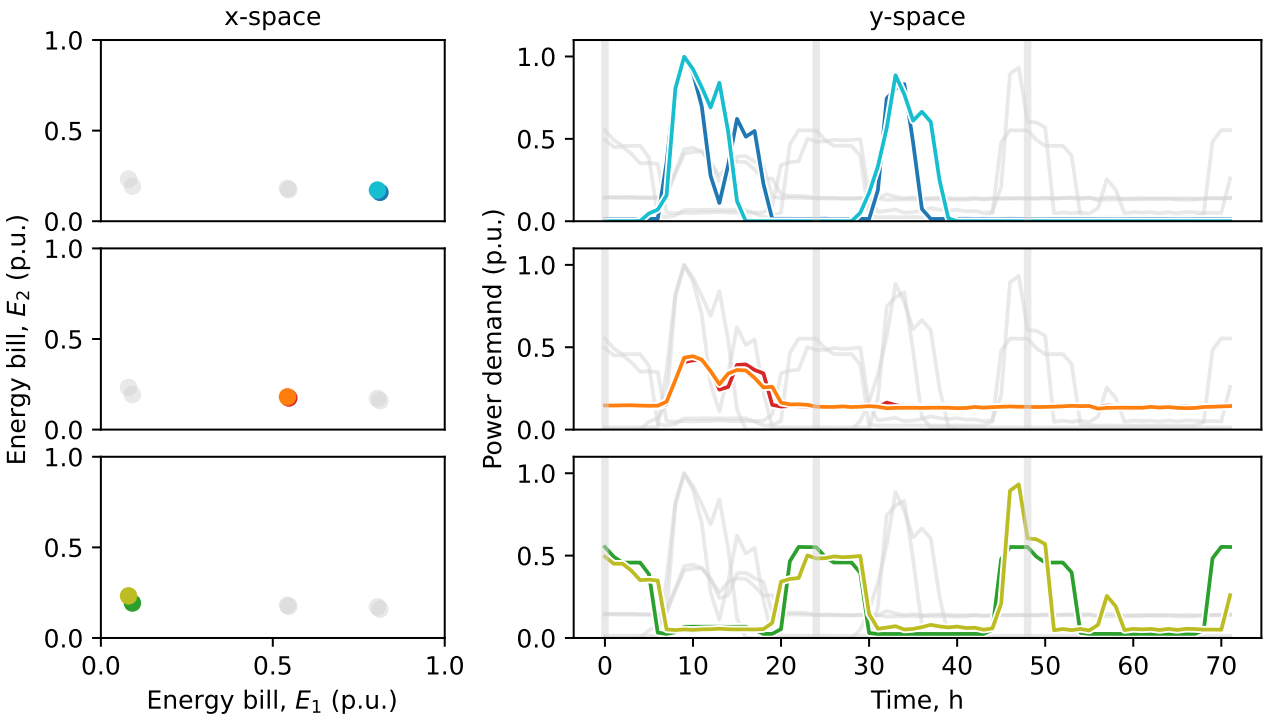

FIGURE 5. Visual examples of the similarity between monthly ToU energy bills and typical hourly load profiles. In particular, three pairs of points (i.e., energy bills) are shown, which are pair-wise close in the x-space. The respective "lines" in the y-space (i.e., typical load profiles) in turn exhibit similar shapes.

IV. METHODS

A. NOTATION AND BASIC CALCULATIONS

A load profile is represented as a vector:

$$\mathbf{p} = \{\mathrm{P_h}\}_{\mathrm{h}=1,\ldots,\mathrm{N_h}}, \tag{1}$$

whose elements $\mathrm{P_h}$ are the average, constant power demand associated with the energy consumption $\mathrm{E_h}$ in each time step $\mathrm{t_h}$, i.e., $\mathrm{P_h} = \mathrm{E_h}/\Delta\mathrm{t_h}$, where $\Delta\mathrm{t_h}$ is one hour; and $\mathrm{N_h}$ the length of the time series.

Given a month-long load profile, it can be chunked into daily sequences, which can be arranged by day type (work days, Saturdays, holidays). We call $\mathrm{L_j}$ the set of load profiles of all days in the month belonging to day type j. Then, the corresponding average load profile $\bar{\mathbf{p}}_\mathrm{j}$ can be evaluated, as follows:

$$\begin{aligned} \bar{\mathbf{p}}_\mathrm{j} &= \left\{\bar{\mathrm{P}}_{\mathrm{j,h}}\right\}_{\mathrm{h}=1,\ldots,\mathrm{N_h}}, \\ \text{s.t. } \bar{\mathrm{P}}_{\mathrm{j,h}} &= \frac{1}{\mathrm{nd_j}} \sum_{\mathrm{l}\in\mathrm{L_j}} \mathrm{P_{l,h}}. \end{aligned} \tag{2}$$

where $\mathrm{nd_j}$ is the size of the set $\mathrm{L_j}$, i.e., the number of days in the month of the j-th day type.

We call J the set of the three average load profiles in a month. Then, the corresponding typical load profile, $\mathbf{y}$, is represented as follows:

$$\mathbf{y} = \{\mathrm{Y_i}\}_{\mathrm{i}=1,\ldots,\mathrm{N_i}} = \left\{\bar{\mathrm{P}}_{\mathrm{j,h}}\right\}_{\mathrm{h}=1,\ldots,\mathrm{N_h}}^{\mathrm{j}=1,\ldots,|\mathrm{J}|}, \tag{3}$$

Considering $\mathrm{N_h} = 24$ and $|\mathrm{J}| = 3$, the length of the typical load profiles is $\mathrm{N_i} = 72$.

An energy bill is also represented as a vector,

$$\mathbf{x} = \{\mathrm{E_f}\}_{\mathrm{f}=1,\ldots,\mathrm{N_f}}, \tag{4}$$

whose elements $\mathrm{E_f}$ are the monthly energy consumption in the three ToU tariffs F1, F2, F3 ($\mathrm{N_f} = 3$).

Given a typical load profile $\mathbf{y}$ the elements $\mathrm{E_f}$ of the associated energy bill vector $\mathbf{x}$ are calculated as follows:

$$\mathrm{E_f} = \sum_{\mathrm{j}=1}^{|\mathrm{J}|=3} \mathrm{nd_j} \cdot \sum_{\mathrm{h}=1}^{\mathrm{N_h}=24} \bar{\mathrm{P}}_{\mathrm{j,h}} \cdot \delta_{\mathrm{f,j}}(\mathrm{t_h}) \cdot \Delta\mathrm{t_h}, \forall \mathrm{f} = 1,2,3, \tag{5}$$

where $\delta_{\mathrm{f,j}}$ is a binary auxiliary variable, defined as follows:

$$\delta_{\mathrm{f,j}}(\mathrm{t_h}) = \begin{cases} 1, & \text{if } \mathrm{t_h} \text{ in day type j belongs to } \mathrm{F_f} \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

whose values, which depend on the hour and type of the day, can be directly obtained from Fig. 2.

B. TRAINING DATA SET

The proposed approach requires a set of energy bills ($\mathbf{x}$) and corresponding typical load profiles ($\mathbf{y}$) pairs. We call these pairs the "training" data set, in compliance with the conventional naming in the field of machine learning. These $\{\mathbf{x}, \mathbf{y}\}$ pairs can be obtained from an end user's year-long load profile, by:

i. dividing the time series into month-long profiles;
ii. repeating the steps in (2), (3) and (5) for each month.

Hence, twelve pairs (one per month) can be obtained from each end user's year-long load profile. Then, the training data set is organized in rows, each corresponding to one of these pairs.

Data normalization

The training data set contains data from heterogeneous end users, so they must be normalized to remove effects related to the total consumption ("magnitude"). Given a pair of $\mathbf{x}$, $\mathbf{y}$ vectors, the elements of the energy bill are normalized so that they sum to 1. Hence, the elements $\hat{\mathrm{E}}_\mathrm{f}$ of the normalized vector $\hat{\mathbf{x}}$ are evaluated as follows:

$$\hat{\mathrm{E}}_\mathrm{f} = \frac{\mathrm{E}_\mathrm{f}}{\sum_\mathrm{f} \mathrm{E}_\mathrm{f}}. \tag{7}$$

Concerning the load profiles, usually min-max [28] or max normalization (e.g., in [37]) are performed to obtain time series values in the range 0–1, or z-standardization to obtain values with null mean and unitary standard deviation (e.g., [18]). In our case, we want to keep the relationship between the typical load profile and energy bill in (5). Therefore, the elements $\hat{\mathrm{Y}}_\mathrm{i}$ of the normalized typical load profile, $\hat{\mathbf{y}}$, are evaluated as follows:

$$\hat{\mathrm{Y}}_\mathrm{i} = \frac{\mathrm{Y}_\mathrm{i}}{\sum_\mathrm{f} \mathrm{E}_\mathrm{f}}. \tag{8}$$

Hence, the resulting normalized typical load profiles have unitary total consumption in the month.

When evaluating the normalized vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, we perform a "row-wise" normalization of the training data set so that all samples can be compared with each other. Usually, also "column-wise" data normalization is required, to make the different features comparable to each other. This is not the case for the energy bill (and the typical load profile), whose single elements are already comparable with each other.

C. K-NN BASED PREDICTION

The training data set is used to predict the end users' typical load profiles from their monthly bills, by means of a k-NN algorithm. This is a well-known supervised learning method [44]. Unlike other algorithms, it does not have a training phase and it exploits the whole training data set in each prediction. The only parameter of the algorithm is the number of neighbors.

Given the training set of ($\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$) pairs, and a vector $\hat{\mathbf{x}}^*$ whose corresponding $\hat{\mathbf{y}}^*$ is unknown, the algorithm works as follows.

i. Evaluate the distance between $\hat{\mathbf{x}}^*$ and all the $\hat{\mathbf{x}}$ vectors in the training data set. The Euclidean distance is used.
ii. Identify the neighbors, i.e., the elements in the training data set with a smaller distance from $\hat{\mathbf{x}}^*$. We call the sets of the $\hat{\mathbf{x}}$ and of the $\hat{\mathbf{y}}$ vectors of the neighbors, respectively, $\mathrm{K_x}$ and $\mathrm{K_y}$. Their size $|\mathrm{K}|$ is equal to the number of neighbors.
iii. Evaluate the prediction ($\hat{\mathbf{y}}_\mathrm{pred}$), as the element-by-element weighted average of the vectors in $\mathrm{K_y}$, as in (9):

$$\hat{\mathbf{y}}_\mathrm{pred} = \sum_\mathrm{k} \frac{1}{|\mathrm{K}|} \cdot \hat{\mathbf{y}}_\mathrm{k}, \tag{9}$$

where $\hat{\mathbf{y}}_\mathrm{k}$ is the typical load profile of the k-th neighbor.
iv. Set $\hat{\mathbf{y}}^*$ equal to the prediction $\hat{\mathbf{y}}_\mathrm{pred}$.

This is the basic implementation of the algorithm, in which the neighbors have the same (uniform) weights However, neighbors could also have different weights, e.g., inversely proportional to the distance between their $\hat{\mathbf{x}}$ vector and $\hat{\mathbf{x}}^*$ [45].

Fig. 6 provides an outline of the process that leads to the k-NN-based prediction of $\hat{\mathbf{y}}^*$. After this procedure, $\mathbf{y}^*$ is obtained by inverting (8), thus restoring the actual magnitude. However, despite the proximity of the neighbors in the x-space, the "predicted" energy bill, evaluated applying (5) to $\mathbf{y}^*$ may not coincide with the original bill in the different ToU tariffs (see Fig. 6).

D. BENCHMARK APPROACHES

We present a general outline of the benchmark approaches against which we compare the proposed one.

Standard load profiles

This method is based on the SLP evaluated for different categories of end users. In particular, we used the SLP adopted by the GSE in [24], defined for two categories, i.e., household and non-household (see Fig. 7). Given an end user with a monthly bill $\mathbf{x}^*$, the associated typical load profile $\mathbf{y}^*$ is evaluated by:

i. selecting a standard profile $\mathbf{y}_\mathrm{ref}$ according the end-user category;
ii. multiplying the SLP by a scaling factor to match the actual total consumption in the month.

Clustering-classification

The detailed steps of this kind of procedure can be found in works like [13], [14] and mainly [15], from which we borrowed this method, therefore we present here only an overview of the main steps and of the algorithms adopted for each task.

The first step is the extraction of the representative load profiles from the training data set, through a clustering performed in the y-space (i.e., only the typical load profiles, $\hat{\mathbf{y}}$, are used). We use K-means, a well-known unsupervised learning algorithm that has been recently proven particularly effective in load profiles clustering [34]. The number of clusters is the only parameter of the algorithm. Based on the result of the clustering, the ($\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$) pairs in the training data set are divided into classes (i.e., the clusters), each with a representative load profile (i.e., the cluster center). Then, a classifier system can be trained using the elements of

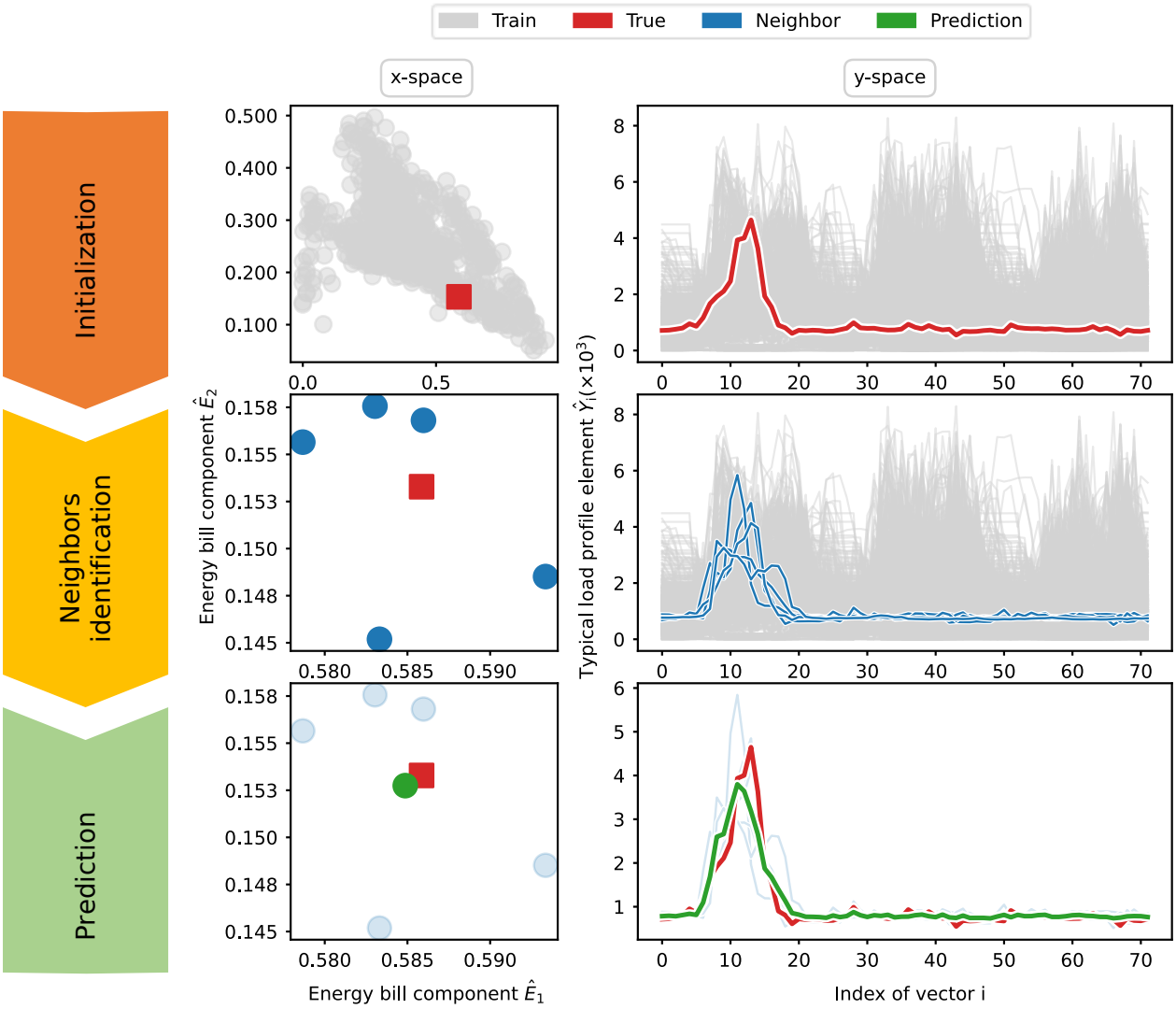

FIGURE 6. Graphical outline of the k-NN prediction process in the x and y-spaces (note that only two components of the bill are shown, to allow a two-dimensional representation of the x-space.). First, the training set composed of ($\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$) pairs is initialized. A test pair of $\hat{\mathbf{x}}^*$ and $\hat{\mathbf{y}}^*$ vectors is also shown (normally $\hat{\mathbf{y}}^*$ is unknown). To predict $\hat{\mathbf{y}}_{\text{pred}}$ from $\hat{\mathbf{x}}^*$, the k nearest neighbors (in the x-space) are evaluated and selected. Then, $\hat{\mathbf{y}}_{\text{pred}}$ is evaluated as the element-by-element weighted average of the neighbors (in the y-space).

the energy bills $\hat{\mathbf{x}}$ as features and the cluster label as targets. This supervised learning task is performed by means of a Decision Tree classifier, as in [15]. At the end of this procedure, a set of rules have been derived, based on which an energy bill can be assigned to one of these classes. When a typical load profile $\hat{\mathbf{y}}^*$ must be predicted from an end user's energy bill $\hat{\mathbf{x}}^*$, a class label is assigned based on the derived decision rules. Then, the cluster center is set as the representative load profile and (8) is inverted to add the magnitude effect so that the total consumption matches the actual energy bill.

In both previous cases, similarly to the proposed approach, while the total energy consumption matches the value in the energy bill, it is not guaranteed that the proportions between the different ToU components coincide. The latter depend indeed on the shape of the predicted load profile, which is imposed by the SLP/cluster centers.

E. VALIDATION

We validate our result using a test set of measured hourly consumption data. Therefore, we first predict typical load profiles using the described methods and then compare them with the real ones. Kohler et al. [25] provided an extensive review of metrics commonly used to compare predicted (or synthetic) and real load profiles, focusing on the difference between "sameness" and "similarity". The former is the time step-by-time step equivalence between two time series. For instance, the mean squared error (MSE) and the mean absolute error (MAE) measure the sameness [25]. The similarity is a broader (and looser) concept that can be declined

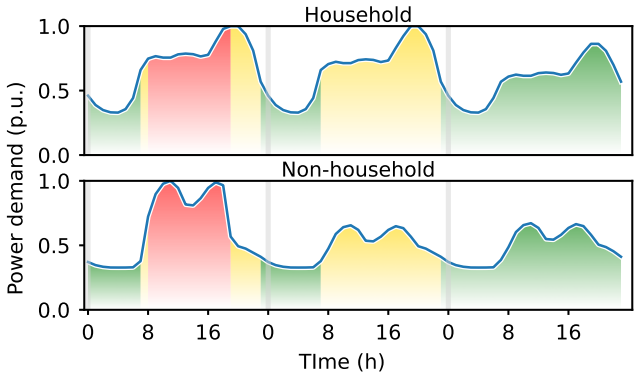

FIGURE 7. **Typical load profiles evaluated from the SLPs adopted by the GSE for household and non-household end users. The subdivision of the time steps into ToU tariffs is also shown (in red, F1, yellow, F2, green, F3). [Elaboration of data from [24]].**

and assessed in different ways. Many metrics proposed in [25] to assess similarity focus on the statistics of the real and simulated profiles (e.g., minimum, median, and maximum values, standard deviation, and error on the duration curve). In the paper, they also propose metrics based on complexity, such as the number of peaks and the fractal dimension.

In this paper, we evaluate the normalized MAE (NMAE) and the Pearson coefficient (r) between the predicted and real load profiles to assess the sameness. We prefer the MAE over other error metrics, such as the MSE, because it directly quantifies the energy consumption that is allocated in the wrong time steps [42]. The r coefficient instead evaluates the linear correlation between two quantities. We also evaluate the ability of the methods to reconstruct the statistics of the data by measuring the NMAE between the duration curves of the predicted and real typical load profiles [25]. We call this metric the duration curve error (DCE).

In all cases, even if we use an artificial profile of 72 hours, we consider the weight of each day type in one month (e.g. there are more work days than Saturdays). As to the NMAE, we perform a weighted average (on the number of days of each day type) of the metric evaluated on the single day type. In the case of the DCE, instead, we evaluate the duration curves of the equivalent month-long load profile by assigning to each day the average load profile of the related day type.

All methods guarantee that the actual total consumption in a month is respected by properly scaling the typical load profile. However, it is not guaranteed that the proportions between the consumption in the different ToU tariffs match the actual one. Therefore, we also evaluate the NMAE between the real and predicted energy bills. Similarly to the case of the load profiles, this metric measures how much of the total monthly consumption is allocated in the wrong tariff by the prediction (more details are reported in Section VI).

V. CASE STUDY

We tested the proposed and benchmark approaches on a data set consisting of 114 end users of different types (i.e., household, DOM, and non-household, BTA), classes (i.e., levels of contractual power, see Table 1), and categories (see Fig. 8). The data were provided by a local energy supplier, who measured the hourly consumption for these end users over a time span of one year. The samples are mostly uniformly distributed among classes (around twelve end users each) and categories. The only exception is represented by the class DOM1, which is composed of only one user. For this reason, we created the class DOM12 by merging DOM1 into DOM2.

TABLE 1. **Codes assigned to the end-user classes according to the type (household, DOM, and non-household, BTA) and level of contractual power at the point-of-delivery defined by the Italian Regulatory Authority for Energy, Networks and the Environment.**

	1	2	3	4	5	6
DOM	≤1.5	1.5–3	>3	∀[1]	-	-
BTA	≤1.5	1.5–3	3–6	6–10	10–16.5	>16.5

[1] Non-resident

Fig. 9 shows the statistics of the consumption in the typical load profiles of the end users in all months, divided by classes. In some cases, consumption patterns with good intra-cluster properties can be found in these classes. For instance, in BTA1, BTA4, BTA5, DOM3, and DOM4 the interquartile range (IQR) appears quite narrow and it follows the median value, while in other cases (e.g., BTA2) it is rather wide. However, even when the IQR is narrow, the extreme values (min, max) can be far from the median and also have different shapes (e.g., BTA1 and BTA4). Finally, Fig. 10 shows the distribution of the three components of the energy bills in the end users, divided by classes. Again, there are cases (e.g., household end users) where the distributions are narrow, but also cases in which they are very dispersed, meaning that end users with very different energy bill composition (hence, consumption patterns) can be found in the same class.

The typical profiles and related energy bills were evaluated from raw data, therefore they have been analyzed in order to eliminate samples with evident inconsistencies or extreme outliers (e.g., time steps in which the consumption was more than five times higher than the average one). This procedure removed about 5% of the original samples: from 114 yearly load profiles, 1368 pairs could be obtained, and 1294 remained after data cleaning.

We used the data set for both testing and training the data-driven approach. Therefore, in order to assess the performance, we reconstructed the data set through a leave-one-out cross-validation [46], where the samples of one end user at a time served as testing while those of

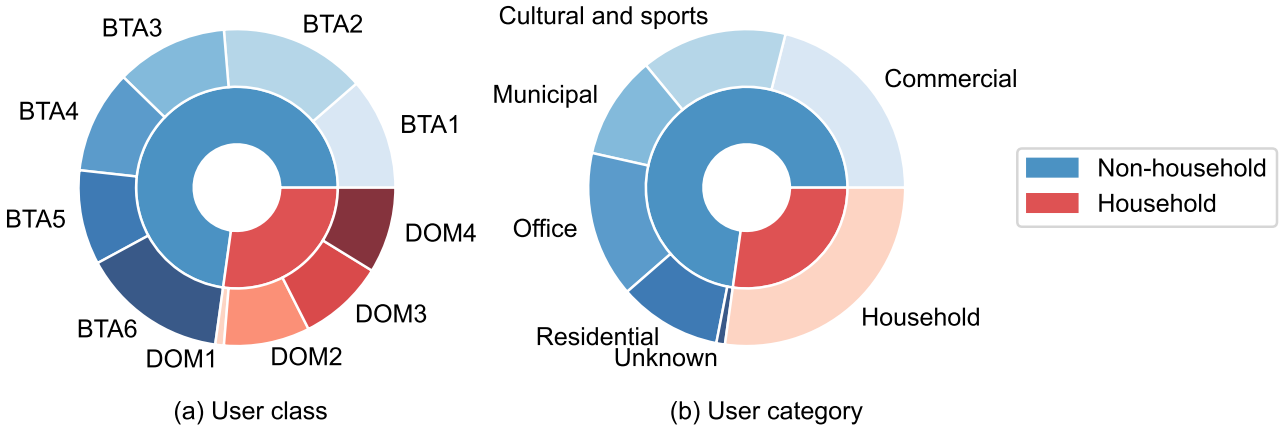

FIGURE 8. Composition of the data set by types of end users (non-household, BTA, and household, DOM) and: (a) class, i.e., level of contractual power at the point-of-delivery; (b) category.

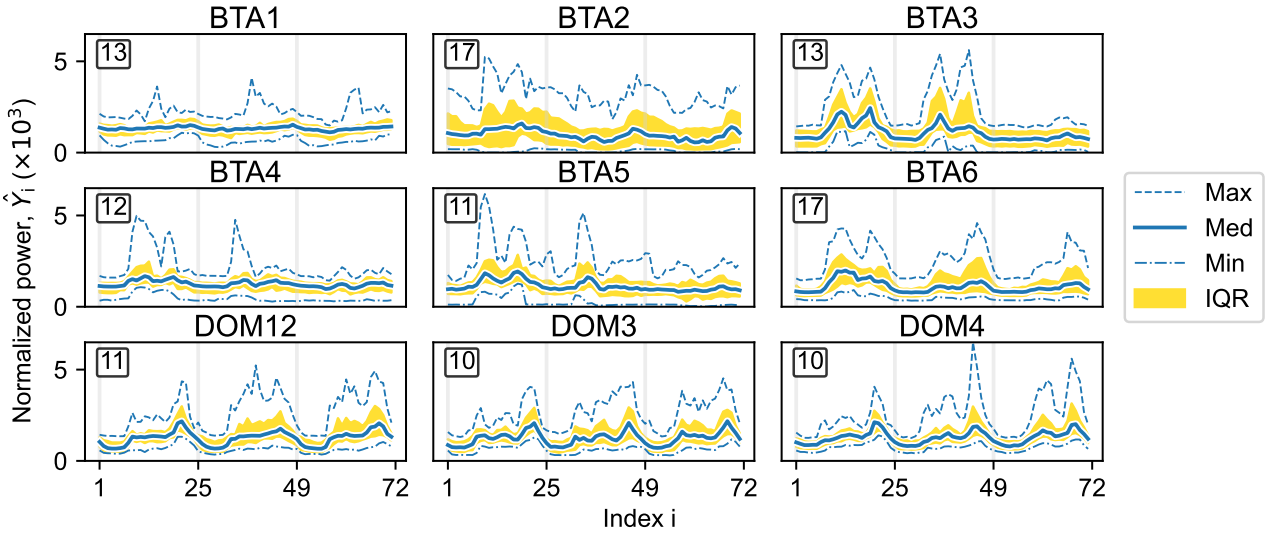

FIGURE 9. Statistics of the typical load profiles evaluated for the end users divided by class: minimum (i.e., 5th percentile), median and maximum(i.e., 95th percentile), and interquartile range, IQR, (25th-75th percentile) of the consumption in each time step. The box in the top-right corner of the plots reports the total number of end users for each class.

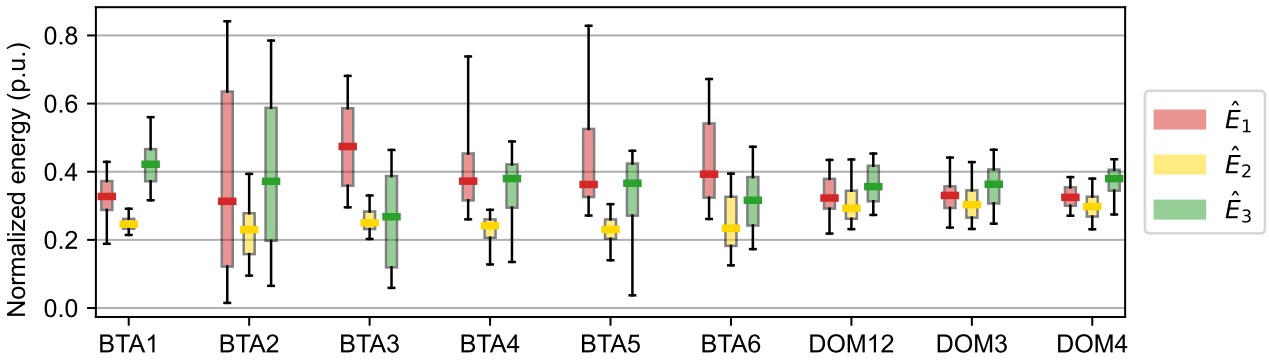

FIGURE 10. Statistics of the monthly energy bills evaluated for the end users divided by ToU tariff time slot and class. The whiskers are truncated at the 5th and 95th percentile.

the other end users served as training data. Accordingly, the number of folds in the cross-validation procedure is equal to the total number of end users in the data set (i.e., 114, each consisting of around 12 profiles).

We implemented all the methods in Python and exploited the open-source library `sklearn` [47] for the K-means, Decision Tree classifier, and k-NN algorithms. We identified a number of clusters equal to 11 (for the clustering-classification approach), and a number of neighbors equal to 9 (for the proposed k-NN) as those minimizing the average error on the predicted load profiles, while we used the default settings from `sklearn` for the training of the Decision Tree.

VI. RESULTS AND DISCUSSION

Fig. 11 shows the distributions of the NMAE between the real and predicted data for the three methods tested. In particular, Fig. 11a shows the NMAE on the ToU components of the energy bill. Thanks to the procedure based on the proximity in the x-space, the k-NN method obtains the smallest error. In the case of the clustering-classification approach, the composition of the predicted energy bill is more "rigid" since it depends on the shape of the cluster centers. This effect is more pronounced in the SLP approach, where only two shapes are possible (household and non-household). Fig. 11b shows the same error metric evaluated on the typical load profiles. In this case, the errors are larger since they are evaluated instantaneously time step by time step. Also in this case, the k-NN shows the best performance, allowing for an error reduction, on average, of 6.5% and 13.8% if compared, respectively, to the clustering-classification and to the SLP approaches.

Fig. 12 shows two different metrics evaluated on the reconstructed yearly load profiles of each end user. Fig. 12a reports the error between the real and predicted duration curves (DCE), while Fig. 12b shows the correlation coefficient between the two load profiles. The former measures the sameness of the statistics of the real and predicted load profiles. Therefore, the DCE is smaller than the NMAE on the typical load profiles. The correlation coefficient instead measures how the real and predicted profiles are linearly correlated. The k-NN approach shows the closest values to 1, which means perfect linear correlation. However, in certain cases, the values of r are significantly smaller than 1.

The radar plots in Fig. 13 show, respectively, the NMAE and r metrics between the real and predicted typical load profiles, divided by end-user classes. In the case of household end users, the performances of the SLP approach are comparable to those of the data-driven ones. However, the SLP reveals to be inadequate to characterize the load profiles of non-household end users, where instead the data-driven approaches, in particular the k-NN, show significantly better performance, comparable to the ones they obtain with the household end users. This means that an SLP is able to characterize the typical load profiles of household end users, where lower within-class variability is found (see Fig. 9). On the other hand, non-residential customers have more diversified consumption patterns, which a unique SLP fails to represent, while data-driven methods can better identify the different shapes based on the energy bill.

The radar plots in Fig. 14, where the NMAE and r are divided by day type, show that all methods perform significantly better on work days than on Saturdays and Sundays/holidays. This can be related to the fact that the consumption in work days is mostly under the ToU tariff F1, which does not appear in other day types. On the contrary, Sundays are completely in ToU tariff F3, which also belongs to night hours of work days and Saturdays therefore, it is more difficult to properly divide the consumption in F3 between the correct time steps.

Discussion

The general trend that emerged from the previous results is that data-driven methods can increase the performances of the prediction of typical load profiles from the ToU energy bill, with respect to SLPs. In fact, they can map the similarity between bills (x) to the similarity between consumption patterns (y). This is particularly true for non-residential end users. Between the two ways of tackling the problem, i.e., the clustering-classification approach that creates a discrete mapping between the x and y-spaces, and the k-NN approach that tends to create a continuous mapping, the latter shows the best performances. As to the different metrics analyzed, worse performances are obtained when comparing the real and predicted data time step by time step, as opposed to the comparison between the statistics of the consumption or the "coarse" granularity data of the ToU bills. Concerning the latter comparison, it should be noted that more sophisticated scaling procedures could be implemented to guarantee that the predicted consumption in each ToU tariff matches the actual one (see for instance [48]). However, these procedures scale the profile with a different factor for each ToU tariff, and therefore they can introduce a distortion of the shape of the predicted load profile. In particular, they can give rise to unrealistic changes in consumption in the hours on the interface between different ToU tariffs.

Limitations and further improvement

The proposed approach allows the prediction of typical load profiles from minimal data that can be easily obtained for a large number of customers, independently from the category of end users. In general, thanks to the k-NN regression, we obtained an average error of less than 4% on the monthly consumption in the different ToU tariffs and of around 26% on the predicted typical load profiles. This means that a quarter of the total consumption is allocated in the wrong time step. The smallest errors are obtained in the prediction of load profiles during work days, which are more frequent than other day types. It is up to the final user of the method

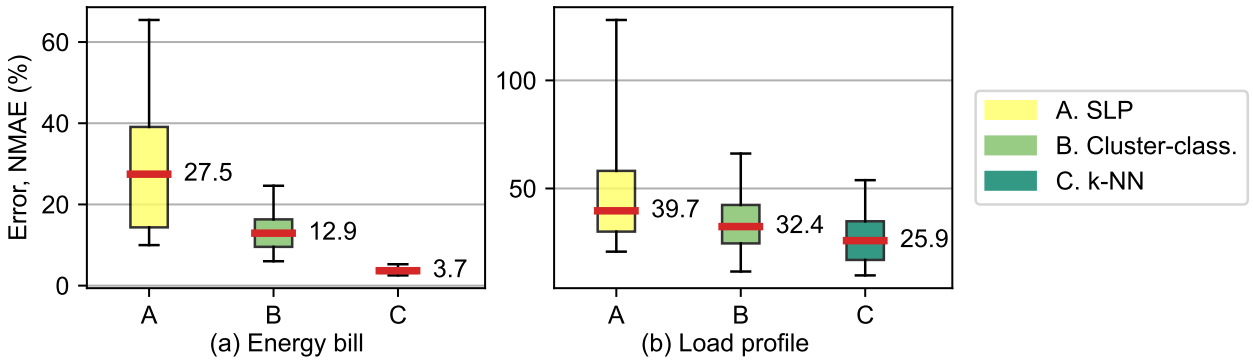

FIGURE 11. Normalized mean absolute error (NMAE) between: (a) the components of the real energy bills and the ones evaluated from the predicted typical load profiles; (b) the real and predicted typical load profiles (weighted on the number of days of each day types in the month), for all samples and the different methods. The median error is shown (red line and value), while the whiskers are truncated at the 5th and 95th percentile.

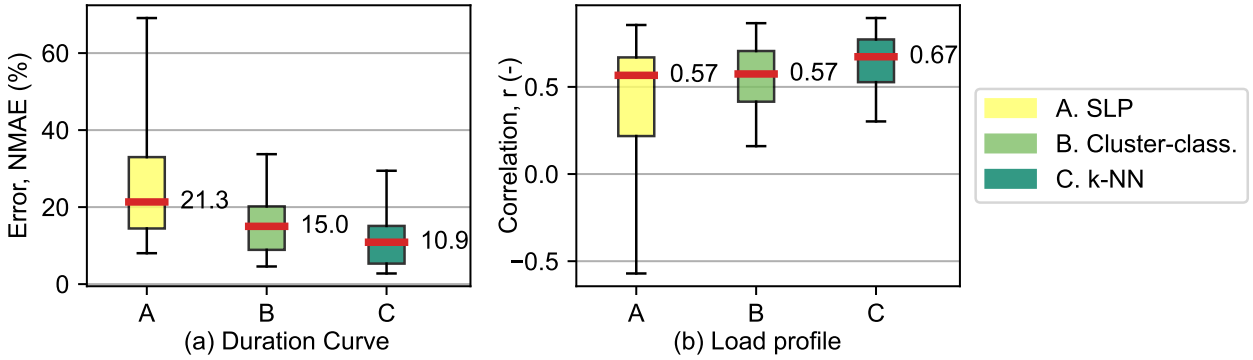

FIGURE 12. (a) Duration curve error (DCE) and (b) Pearson correlation coefficient (r) between the real and predicted reconstructed yearly load profiles, for all the samples and the different methods. The median error of each method is shown (red line and value), while the whiskers are truncated at the 5th and 95th percentile.

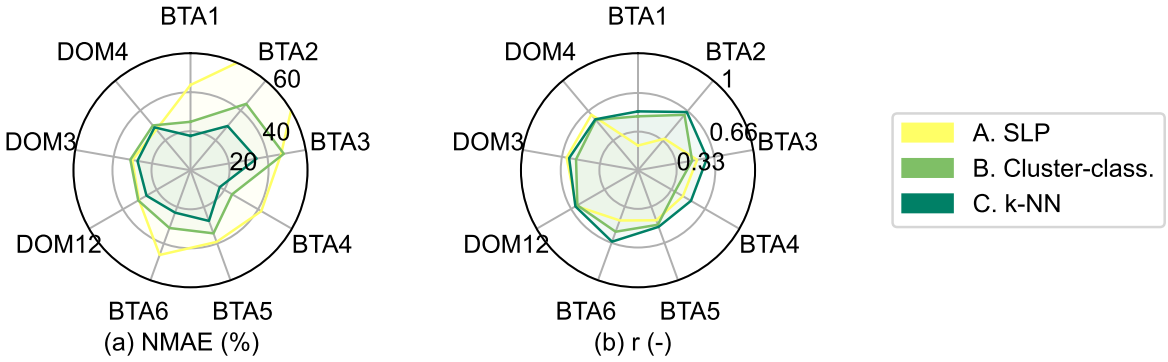

FIGURE 13. Median value of: (a) Normalized mean absolute error (NMAE) and (b) Pearson correlation coefficient (r) between the real and predicted load profiles, for all the samples and the different methods, broken down by end-user categories.

TABLE 2. Median value (and interquartile range) of NMAE on typical load profiles between the real and predicted data, broken down by end user categories, obtained with the different methods: A. Standard Load Profile; B. Clustering-classification; C. k-NN.

Method	BTA1	BTA2	BTA3	BTA4	BTA5	BTA6	DOM12	DOM3	DOM4
A	43.7 (19.4)	95.2 (68.5)	47.1 (36.8)	41.9 (18.9)	39.1 (28.6)	46.2 (22.8)	30.5 (9.0)	30.3 (10.6)	27.2 (9.4)
B	24.8 (19.4)	44.4 (28.2)	48.6 (26.2)	24.6 (14.1)	34.4 (17.7)	31.5 (14.7)	30.9 (7.3)	31.3 (11.3)	30.0 (6.3)
C	17.6 (6.8)	29.6 (18.9)	34.4 (20.0)	17.3 (6.9)	27.6 (15.4)	23.1 (10.8)	26.0 (7.0)	27.6 (5.8)	28.7 (6.8)

to decide whether the compromise between accuracy and ease of data collection is acceptable. However, the proposed approach leaves room for further improvement. In particular, given the size of our test data set, we

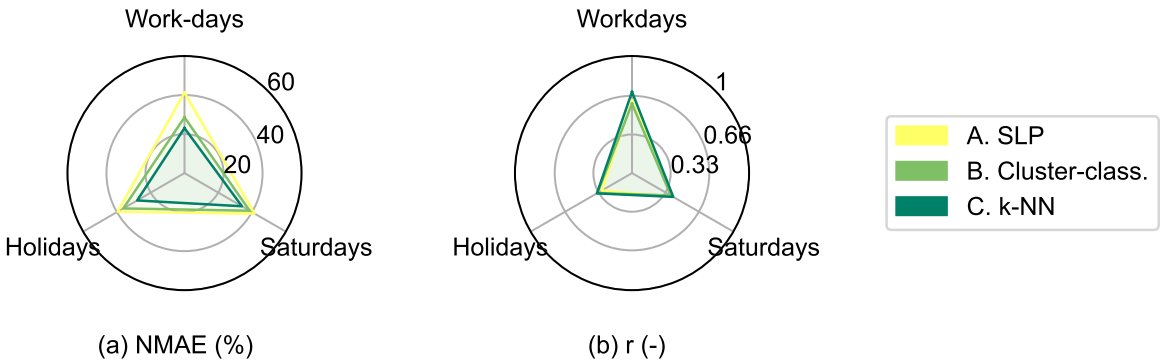

FIGURE 14. Median value of: (a) Normalized mean absolute error (NMAE) and (b) Pearson correlation coefficient (r) between the real and predicted load profiles, for all the samples and the different methods, broken down by day type.

TABLE 3. Median value (and interquartile range) of NMAE on duration yearly reconstructed duration curve (duration curve error, DCE) between the real and predicted data, broken down by end user categories, obtained with the different methods: A. Standard Load Profile; B. Clustering-classification; C. k-NN.

Method	BTA1	BTA2	BTA3	BTA4	BTA5	BTA6	DOM12	DOM3	DOM4
A	21.0 (12.4)	48.9 (35.4)	31.1 (33.0)	24.2 (13.8)	18.8 (10.2)	16.2 (13.2)	19.1 (10.1)	18.7 (19.3)	16.6 (10.5)
B	10.3 (13.9)	24.0 (15.6)	17.2 (8.4)	8.2 (8.1)	13.3 (6.7)	17.4 (10.3)	15.4 (6.4)	13.8 (13.7)	15.0 (5.3)
C	5.5 (9.2)	13.0 (8.8)	12.7 (14.9)	4.0 (8.2)	13.0 (8.5)	11.3 (9.4)	11.3 (4.3)	10.7 (6.0)	11.3 (2.5)

TABLE 4. Median value (and interquartile range) of NMAE on ToU energy bills between the real and predicted data, broken down by end user categories, obtained with the different methods: A. Standard Load Profile; B. Clustering-classification; C. k-NN.

Method	BTA1	BTA2	BTA3	BTA4	BTA5	BTA6	DOM12	DOM3	DOM4
A	36.9 (17.9)	49.4 (46.0)	32.5 (16.0)	31.3 (16.7)	33.7 (14.7)	28.9 (14.2)	14.2 (4.8)	13.6 (6.2)	11.5 (2.7)
B	10.6 (5.5)	15.3 (11.1)	15.7 (6.4)	13.7 (8.0)	15.3 (9.8)	14.3 (4.9)	10.8 (5.5)	9.7 (5.1)	11.0 (4.8)
C	3.6 (0.7)	3.8 (1.5)	4.2 (1.2)	3.3 (0.6)	3.3 (0.5)	3.9 (2.1)	3.8 (0.4)	3.8 (0.3)	3.9 (0.6)

deployed a single model for all end-user categories and all months/seasons. Different k-NN regressions could be used to predict the typical load profiles of end users from different categories/seasons when a richer data set is available (so to cover uniformly the spaces of the energy bills and typical load profiles). It should be noted that the further loss of detail related to the use of typical load profiles is not considered here. Finally, the methods have been tested on a set of end users located in the same geographical area, thus temperature-related effects are compensated by similar weather conditions. It is still to be assessed whether a single model or more models (hence, more data set) are needed to work with end users of different regions. Nonetheless, we believe that the proposed method provides a valid option for the prediction of a large number of different end users' load profiles when few input data are available.

VII. CONCLUSION

In this work, we proposed a data-driven approach to predict load profiles in typical days (work days, Saturdays, holidays) for end users of heterogeneous categories, and from few and easy-to-collect input data. In particular, we proposed a method that predicts typical load profiles based on the similarity between monthly time-of-use energy bills, using a k-nearest neighbors algorithm. We assessed the performances of the proposed method in comparison to two benchmarks: an approach based on Standard Load Profiles and a data-driven method based on the identification of similar load profiles (clustering) and of decision rules to assign new customers to one cluster, hence to its representative load profile, based on their time-of-use energy bill (classification).

All methods allow the prediction of typical load profiles of end user of different categories from a few input data (i.e., only the monthly bill). The results obtained on a data set of measured hourly consumption show that the methods have poorer performances on error metrics evaluating the sameness of the predicted load profiles to the real ones, while better performances are achieved in the error metrics that measure the statistics of the consumption (e.g., duration curve). The results also highlight that the proposed method outperforms the other ones basically in all error metrics and for each end-user category (with few exceptions).

In future works on the load profile prediction from electricity bills, the analysis of the results can be deepened both in terms of diversity of the similarity/sameness metrics and in terms of benchmark meth-

ods to further assess where the proposed method stands in terms of the trade-off between accuracy and ease of data collection. Furthermore, the adoption of multiple k-NN models to be used for different end-user categories and/or different seasons can be explored, if richer data set are available. Furthermore, it should be noticed that in this work we preferred to utilize well-known and easily-interpretable methods for our analysis. However, different and more advanced methods from the field of machine learning [38] could be explored to better assess the potentiality of the proposed approach.

DATA AND CODE AVAILABILITY

The processed data set used for training and testing the proposed and benchmark approaches is available at: https://github.com/cadema-PoliTO/Bill2Watt under CC-BY-NC 4.0 License. Upon publication, the code will eventually be made available under CC-BY-4.0 License at the same address.

ACKNOWLEDGMENT

The authors would like to express their gratitude towards Ezio Chiaramello and Eleonora Barbuzza of Acea Pinerolese Energia for providing the data.

REFERENCES

[1] R. Masiello, R. Fioravanti, B. Chalamala, and H. Passell, "Electrification, decarbonization, and the future carbon-free grid: the role of energy storage in the electric grid infrastructure [point of view]," Proceedings of the IEEE, vol. 110, no. 3, pp. 324–333, 2022, [Crossref].

[2] R. Haas, N. Duic, H. Auer, A. Ajanovic, J. Ramsebner, J. Knapek, and S. Zwickl-Bernhard, "The photovoltaic revolution is on: How it will change the electricity system in a lasting way," Energy, vol. 265, p. 126351, 2023, [Crossref].

[3] P. Lazzeroni, S. Olivero, and M. Repetto, "Economic perspective for PV under new Italian regulatory framework," Renewable and Sustainable Energy Reviews, vol. 71, pp. 283–295, 2017, [Crossref].

[4] A. Canova, P. Lazzeroni, G. Lorenti, F. Moraglio, A. Porcelli, and M. Repetto, "Decarbonizing residential energy consumption under the Italian collective self-consumption regulation," Sustainable Cities and Society, vol. 87, p. 104196, 2022, [Crossref].

[5] T. Beck, H. Kondziella, G. Huard, and T. Bruckner, "Assessing the influence of the temporal resolution of electrical load and PV generation profiles on self-consumption and sizing of PV-battery systems," Applied Energy, vol. 173, pp. 331–342, 2016, [Crossref].

[6] A. Cielo, P. Margiaria, P. Lazzeroni, I. Mariuzzo, and M. Repetto, "Renewable Energy Communities business models under the 2020 Italian regulation," Journal of Cleaner Production, vol. 316, p. 128217, 2021, [Crossref].

[7] H. Kazmi, Íngrid Munné-Collado, F. Mehmood, T. A. Syed, and J. Driesen, "Towards data-driven energy communities: A review of open-source datasets, models and tools," Renewable and Sustainable Energy Reviews, vol. 148, p. 111290, 2021, [Crossref].

[8] J. Coignard, "Energy Communities: Sharing Resources on the Distribution Grid," Ph.D. dissertation, Université Grenoble Alpes, Grenoble, France, 2022, (accessed Apr. 05, 2023).

[9] A. Piti, D. Mardero, A. Signorini, A. Boscagnin, G. Ceneri, and A. Cammarota, "Smart Metering 2G – Evolution of a Smart Metering Experience," in 25th International Conference on Electricity Distribution (CIRED), Spain: Madrid, 2019.

[10] S. Vitiello, N. Andreadou, M. Ardelean, and G. Fulli, "Smart metering roll-out in europe: Where do we stand? cost benefit analyses in the clean energy package and research trends in the green deal," Energies, vol. 15, p. 2340, 2022, [Crossref].

[11] C. Stegner, O. Glaß, and T. Beikircher, "Comparing smart metered, residential power demand with standard load profiles," Sustainable Energy, Grids and Networks, vol. 20, p. 100248, 2019, [Crossref].

[12] E. Proedrou, "A comprehensive review of residential electricity load profile models," IEEE Access, vol. 9, pp. 12 114–12 133, 2021, [Crossref].

[13] D. Vercamer, B. Steurtewagen, D. Van den Poel, and F. Vermeulen, "Predicting consumer load profiles using commercial and open data," IEEE Transactions on Power Systems, vol. 31, no. 5, pp. 3693–3701, 2016, [Crossref].

[14] J. L. Viegas, S. M. Vieira, R. Melício, V. Mendes, and J. M. Sousa, "Classification of new electricity customers based on surveys and smart metering data," Energy, vol. 107, pp. 804–817, 2016, [Crossref].

[15] M. S. Piscitelli, S. Brandi, and A. Capozzoli, "Recognition and classification of typical load profiles in buildings with non-intrusive learning approach," Applied Energy, vol. 255, p. 113727, 2019, [Crossref].

[16] ACER/CEER, "Annual Report on the Results of Monitoring the Internal Electricity and Gas Markets in 2015," https://www.acer.europa.eu/, 2016, (accessed Jan. 20, 2023).

[17] IRENA, International Renewable Energy Agency, "Innovation landscape brief: Time-of-use tariffs," https://www.irena.org/, 2019, (accessed Jan. 20, 2023).

[18] J. Y. Park, X. Yang, C. Miller, P. Arjunan, and Z. Nagy, "Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset," Applied Energy, vol. 236, pp. 1280–1295, 2019, [Crossref].

[19] X. Kang, J. An, and D. Yan, "A systematic review of building electricity use profile models," Energy and Buildings, vol. 281, p. 112753, 2023, [Crossref].

[20] T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas, and M. Kolehmainen, "Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data," Applied Energy, vol. 87, no. 11, pp. 3538–3545, 2010, [Crossref].

[21] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," Renewable and Sustainable Energy Reviews, vol. 120, p. 109628, 2020, [Crossref].

[22] S. Ramos, J. M. Duarte, F. J. Duarte, and Z. Vale, "A data-mining-based methodology to support mv electricity customers' characterization," Energy and Buildings, vol. 91, pp. 16–25, 2015, [Crossref].

[23] M. Anvari, E. Proedrou, B. Schäfer, C. Beck, H. Kantz, and M. Timme, "Data-driven load profiles and the dynamics of residential electricity consumption," Nature Communications, vol. 13, p. 4593, 2022, [Crossref].

[24] GSE, "Modalità di profilazione dei dati di misura e relative modalità di utilizzo [Modalities for profiling measured data and related usage]." https://www.gse.it/, 2022, (accessed Jan. 20, 2023).

[25] S. Köhler, R. Rongstock, M. Hein, and U. Eicker, "Similarity measures and comparison methods for residential electricity load profiles," Energy and Buildings, vol. 271, p. 112327, 2022, [Crossref].

[26] A. Grandjean, J. Adnot, and G. Binet, "A review and an analysis of the residential electric load curve models," Renewable and Sustainable Energy Reviews, vol. 16, no. 9, pp. 6539–6565, 2012, [Crossref].

[27] E. M. S. Duque, P. P. Vergara, P. H. Nguyen, A. van der Molen, and J. G. Slootweg, "Conditional multivariate elliptical copulas to model residential load profiles from smart meter data," IEEE Transactions on Smart Grid, vol. 12, no. 5, pp. 4280–4294, 2021, [Crossref].

[28] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern

grouping," Energy, vol. 42, no. 1, pp. 68–80, 2012, https://doi.org/10.1016/j.energy.2011.12.031.
[29] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," IEEE Transactions on Smart Grid, vol. 10, no. 3, pp. 3125–3148, 2019, [Crossref].
[30] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, and J. Li, "A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis," Energy and Built Environment, vol. 1, no. 2, pp. 149–164, 2020, [Crossref].
[31] T.-c. Fu, "A review on time series data mining," Engineering Applications of Artificial Intelligence, vol. 24, no. 1, pp. 164–181, 2011, https://doi.org/10.1016/j.engappai.2010.09.007.
[32] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour, "Energy Forecasting: A Review and Outlook," IEEE Open Access Journal of Power and Energy, vol. 7, pp. 376–388, 2020, [Crossref].
[33] X. Luo, T. Hong, Y. Chen, and M. A. Piette, "Electric load shape benchmarking for small- and medium-sized commercial buildings," Applied Energy, vol. 204, pp. 715–725, 2017, [Crossref].
[34] S. Eiraudo, L. Barbierato, R. Giannantonio, A. Porta, A. Lanzini, R. Borchiellini, E. Macii, E. Patti, and L. Bottaccioli, "A machine learning based methodology for load profiles clustering and non-residential buildings benchmarking," IEEE Transactions on Industry Applications, pp. 1–11, 2023, [Crossref].
[35] F. D. Minuto, P. Lazzeroni, R. Borchiellini, S. Olivero, L. Bottaccioli, and A. Lanzini, "Modeling technology retrofit scenarios for the conversion of condominium into an energy community: An Italian case study," Journal of Cleaner Production, vol. 282, p. 124536, 2021, [Crossref].
[36] A. Mutanen, M. Ruska, S. Repo, and P. Jarventausta, "Customer Classification and Load Profiling Method for Distribution Systems," IEEE Transactions on Power Delivery, vol. 26, no. 3, pp. 1755–1763, 2011, [Crossref].
[37] S. Zhan, Z. Liu, A. Chong, and D. Yan, "Building categorization revisited: A clustering-based approach to using smart meter data for building energy benchmarking," Applied Energy, vol. 269, p. 114920, 2020, [Crossref].
[38] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics). Berlin, Heidelberg: Springer-Verlag, 2006.
[39] K. le Zhou, S. lin Yang, and C. Shen, "A review of electric load classification in smart grid environment," Renewable and Sustainable Energy Reviews, vol. 24, pp. 103–110, 2013, [Crossref].
[40] A. Albert and R. Rajagopal, "Smart meter driven segmentation: What your consumption says about you," IEEE Transactions on Power Systems, vol. 28, no. 4, pp. 4019–4030, 2013, [Crossref].
[41] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," Applied Energy, vol. 141, pp. 190–199, 2015, [Crossref].
[42] R. Granell, C. J. Axon, M. Kolokotroni, and D. C. Wallom, "A data-driven approach for electricity load profile prediction of new supermarkets," Energy Procedia, vol. 161, pp. 242–250, 2019, [Crossref].
[43] ARERA, "Proposte in materia di definizione delle fasce orarie per l'anno 2007 e successivi [Proposals for the definition of the time-of-use time slots starting from year 2007.]," https://www.arera.it/, 2006, (accessed Jan. 20, 2023).
[44] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," The American Statistician, vol. 46, no. 3, pp. 175–185, 1992, [Crossref].
[45] S. A. Dudani, "The Distance-Weighted k-Nearest-Neighbor Rule," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-6, no. 4, pp. 325–327, 1976, [Crossref].
[46] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," Pattern Recognition, vol. 48, no. 9, pp. 2839–2846, 2015, [Crossref].
[47] "Scikit-learn: Machine Learning in Python," https://scikit-learn.org/, 2023, (accessed Mar. 14, 2023).
[48] M. Lamagna, B. Nastasi, D. Groppi, M. M. Nezhad, and D. A. Garcia, "Hourly energy profile determination technique from monthly energy bills," Building Simulation, vol. 13, no. 6, pp. 1235–1248, 2020, [Crossref].

PAOLO LAZZERONI received the M.S. degree and the Ph.D. degree in electrical engineering from Politecnico di Torino, Italy, in 2006 and 2011. He is currently an Assistant Professor of electrical engineering with the Energy Department "Galileo Ferraris", Politecnico di Torino, Italy. His main research interests are renewable distributed generation, modeling and optimal management of complex (hybrid) multi-energy systems, electric mobility, multi-objective optimization in energy communities and demand management.

GIANMARCO LORENTI received the M.S. degree in energy and nuclear engineering in 2021, from Politecnico di Torino, Italy, where he is currently a PhD candidate in electrical engineering, with the Energy Department "Galileo Ferraris". His main research activity is related to the modeling, design, and management of complex distributed multi-energy systems through optimization and computational intelligence in the context of collective energy consumption.

MAURIZIO REPETTO Maurizio Repetto (Genova, Italy, 1960) received his Master and Ph. D. degrees in Electrical Engineering from the University of Genova. Since year 2000 is full professor of Principle of Electrical Engineering at the Department of Energy "Galileo Ferraris" of the Politecnico di Torino. From 2013 to 2017, Honorary Professor in the School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia. In 2018 visiting researcher at the University of Ontario Institute of Technology in Oshawa, Canada. Since October 2018 he is elected member for Europe in the Board of the International Compumag Society. In February 2019 has been selected as expert for the Horizontal Working Group (HWG) on 100% Renewable Energy Districts of the of the Renewable Heating and Cooling (RHC) European Platform. Since July 2020, designated as Alternate Representative for Italy in the Executive Committee of the International Energy Agency Implementing Agreement on District Heating & Cooling, including Combined Heat and Power (IEA DHC-TCP). He is author of 233 publications and of five patents. Bibliometrics, listed in the major research databases (Scopus – accessed Mar 4th, 2022), attribute more than 2600 citations (h-index 27).