POLITECNICO DI TORINO
Repository ISTITUZIONALE

Reconstructing hourly residential electrical load profiles for Renewable Energy Communities using non-intrusive machine learning techniques

*Original*
Reconstructing hourly residential electrical load profiles for Renewable Energy Communities using non-intrusive machine learning techniques / Giannuzzo, L., Minuto, F.D., Schiera, D.S., Lanzini, A.. - In: ENERGY AND AI. - ISSN 2666-5468. - ELETTRONICO. - 15:(2024). [10.1016/j.egyai.2023.100329]

*Availability:*
This version is available at: 11583/2984589 since: 2023-12-18T11:23:59Z

*Publisher:*
Elsevier

*Published*
DOI:10.1016/j.egyai.2023.100329

*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)

25 November 2026

Energy and AI 15 (2024) 100329

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.sciencedirect.com/journal/energy-and-ai

# Reconstructing hourly residential electrical load profiles for Renewable Energy Communities using non-intrusive machine learning techniques

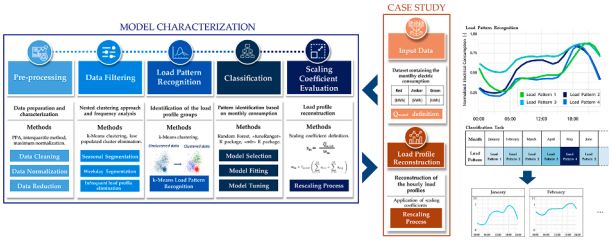
Lorenzo Giannuzzo [a,b,*], Francesco Demetrio Minuto [a,b], Daniele Salvatore Schiera [a,b], Andrea Lanzini [a,b]

[a] Energy Center Lab, Polytechnic of Turin, via Paolo Borsellino 38/16, 10152, Turin, Italy
[b] Department of Energy (DENERG), Polytechnic of Turin, Corso Duca degli Abruzzi 24, 10129, Turin, Italy

## HIGHLIGHTS

- Development of a non-intrusive machine learning model to reconstruct the aggregated electrical profile of residential users at an hourly resolution level to estimate the shared energy of a Renewable Energy Community.
- Formulation of a rescaling process capable of obtaining electrical load profiles from normalized load curves.
- Optimization of a Random Forest algorithm to identify residential users' consumption based only on monthly electrical consumption.
- The methodology is tested on a public dataset, achieving a Normalized Mean Absolute Error (NMAE) and a Normalized Root Mean Square Error (NRMSE) of 20.04 % and 26,17 % comparing the simulated hourly electrical load profile to the real one, and a NMAE and a NRMSE of 18.34 % and 23.87 % during contemporaneity between energy production and aggregate consumption within the REC.
- A Relative Absolute Error in estimating the shared energy at a monthly (MRAE) and yearly (RAE) resolution level is obtained equal to 8.31 % and 0.12 %, respectively.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Keywords:*
Renewable Energy Community
Load profiling

## ABSTRACT

The successful implementation of Renewable Energy Communities (RECs) involves maximizing the self-consumption within a community, particularly in regulatory contexts in which shared energy is incentivized. In many countries, the absence of a metering infrastructure that provides data at an hourly or sub-hourly res-

* Corresponding author at: Via Paolo Braccini 29, 10141, Turin, Italy.
*E-mail address:* lorenzo.giannuzzo@polito.it (L. Giannuzzo).

https://doi.org/10.1016/j.egyai.2023.100329

Available online 9 December 2023
2666-5468/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Non-intrusive machine learning
Data-driven models
Data analytics
Shared energy estimation

olution level for low-voltage users (e.g., residential and commercial users) makes the design of a new energy community a challenging task. This study proposes a non-intrusive machine learning methodology that can be used to generate residential electrical consumption profiles at an hourly resolution level using only monthly consumption data (i.e., billed energy), with the aim of estimating the energy shared by RECs. The proposed methodology involves three phases: first, identifying the typical load patterns of residential users through k-Means clustering, then implementing a Random Forest algorithm, based on monthly energy bills, to identify typical load patterns and, finally, reconstructing the hourly electrical load profile through a data-driven rescaling procedure. The effectiveness of the proposed methodology has been evaluated through an REC case study composed by 37 residential users powered by a 70 $kW_p$ photovoltaic plant. The Normalized Mean Absolute Error (NMAE) and the Normalized Root Mean Squared Error (NRMSE) were evaluated over an entire year and whenever the energy was shared within the REC. The Relative Absolute Error was also measured when estimating the shared energy at both a monthly (MRAE) and at an annual basis. (RAE). A comparison between the REC load profile reconstructed using the proposed methodology and the real load profile yielded an overall NMAE of 20.04 %, an NRMSE of 26.17 %, and errors of 18.34 % and 23.87 % during shared energy timeframes, respectively. Furthermore, our model delivered relative absolute errors for the estimation of the shared energy at a monthly and annual scale of 8.31 % and 0.12 %, respectively.

**Nomenclature**

| | |
|---|---|
| $Q$ | Real Consumption (kWh) |
| $\widehat{Q}$ | Simulated Consumption (kWh) |
| $E$ | Energy (kWh) |
| $T$ | Temporal extension (h) |
| $q$ | Normalized consumption |
| $TV$ | Threshold Value |
| $C$ | Cluster |
| $s$ | Scaling coefficient |
| $w$ | Weight coefficient |
| $NRMSE$ | Normalized Root Mean Square Error |
| $NMAE$ | Normalized Mean Absolute Error |
| $NRAE$ | Normalized Relative Absolute Error |

*Subscripts*

| | |
|---|---|
| $h$ | Hour |
| $m$ | Month |
| $y$ | Year |
| $t$ | Time of Use rate |
| $u$ | User |
| $p$ | Profile |
| $max$ | Maximum |
| $cons$ | Consumption |
| $prod$ | Production |
| $avg$ | Average |
| $SE$ | Shared Energy |

## 1. Introduction

### 1.1. Context and motivation

Renewable Energy Communities (RECs) are grassroots initiatives that invest in clean energy to meet the consumption needs and environmental goals of a community, and thereby contribute to the spread of renewables [1]. RECs are local organizations that bring together citizens, small and medium enterprises (SMEs), institutions, and all the final users to develop projects for the production, sharing, and local use of renewable energy. The clean energy production of RECs generally relies on such technologies as photovoltaic solar panels, wind turbines, and hydroelectric plants, and can include the production of electricity, the heating and cooling of buildings, and the charging of electric vehicles. Moreover, RECs are designed to support the deployment of renewable energy sources [2], promote energy efficiency, the use of renewable sources and the reduction of greenhouse gas emissions, as well as to foster the creation of more sustainable local entities. Additionally, they allow citizens to participate directly in the energy market as prosumers [3], thereby promoting self-consumption and allowing the citizens to

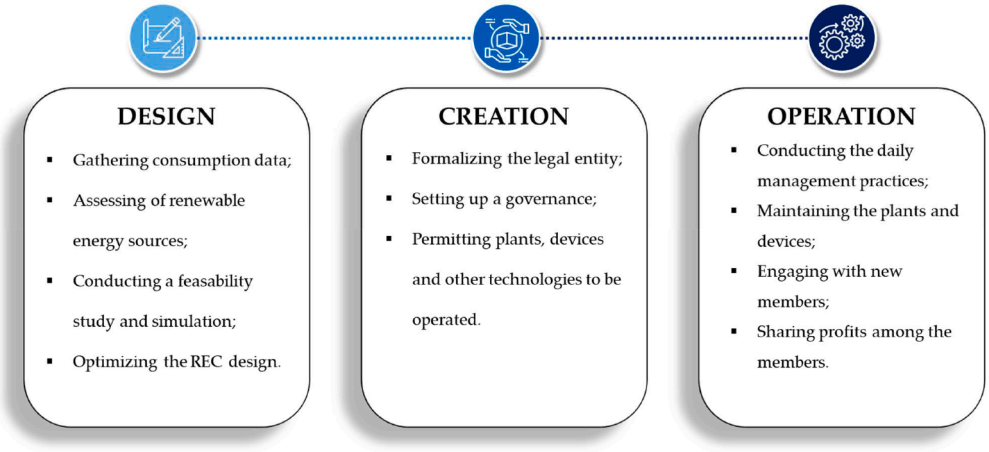



**Fig. 1.** The development timeline of an REC.

become the producers and owners of renewable-powered plants. RECs represent an emerging player on energy markets, where they undertake a variety of activities to deliver benefits to their shareholders and members. They serve as a tangible example of how an energy transition can evolve into a participatory and inclusive phenomenon, thereby enabling the players to become active in the governance of their territory and shaping their energy future [4]. The typical development process of a new REC project is composed of three phases: design, creation, and operation, as depicted in Fig. 1.

The design phase commences with a group of citizens, businesses, institutions, and other local entities that share the same interest and willingness of other members to collaborate in such an endeavor. They first identify the initial core of the project by searching for potential members, verifying their geographical proximity, gathering data on their energy demand, and assessing the availability of renewable energy sources (RES) or the potential for new installations [4]. They then proceed to develop a feasibility study to simulate energy flows, optimize the REC design (i.e., to maximize the shared energy), and formulate a preliminary action plan. Finally, they issue a call to action for all the citizens, investors, and other stakeholders to review and approve the design and investments of the project. The creation phase is initiated by formalizing the legal entity of the Renewable Energy Community, defining the roles of the stakeholders, and establishing the governance structure of the RES. Following this, the necessary funds and permits required to install new generation facilities are obtained. This phase culminates in setting up the generation plant, devices, and technologies necessary to manage and control the REC, and in obtaining the necessary permits to operate such a REC. The operation phase encompasses the daily management of the REC, and includes monitoring the energy flows and performance, maintaining the plants and devices, managing the costs and revenues, engaging with members, providing customer services, and sharing profits among the members.

One of the most significant challenges that can arise during the design phase concerns the availability of hourly or sub-hourly consumption data of the REC members. The availability of such data is closely related to the optimal sizing of the power plants, which is a key aspect for the energetic and economic stability of an REC project, especially in incentive-driven contexts where a major part of the REC income is related to the economic valorization of shared energy. A multitude of research studies have been conducted regarding RECs. Such studies have encompassed various aspects, such as the impact of RECs in complex socio-economic contexts [5–7], the utilization of digital platforms as tools to support their development [4], policies [8,9], and their optimal sizing and design. The most recent studies have examined the aggregation of prosumers within RECs from both an energetic and an economic standpoint [10]. These studies have also taken into consideration communal batteries and reversible solid oxide cells, as well as their impact on the distribution grid [11–13]. Furthermore, novel business models and optimization methodologies which have focused on energy communities that introduce fair revenue sharing policies, the equitable payment of aggregators, and exit clauses have been presented [14,15]. These studies have also proposed innovative models and methodologies that can be used to determine optimal management strategies for energy systems, even in incentive-driven environments [16,17]. Additionally, AI models have been utilized to explore the potential integration of multi-energy and storage systems, as well as their management [18–20].

However, it is important to note that these recent research studies have relied heavily on high-resolution data of real consumptions, which are challenging to obtain during the design phase of an REC, particularly for the residential sector [16]. Furthermore, the models utilized in these studies require such data as those depicted in Table 1.

As can be observed, in the aforementioned cases, the optimization, simulation, sharing mechanism evaluation and REC management strategy analyses were developed using high-resolution data. However, in real-context scenarios, obtaining data at an hourly or sub-hourly resolution level is often challenging, thereby making it difficult to exploit the models utilized in previous research works, such as MILP, Energy Hub, and Time Delay Neural Network.

The availability of hourly or sub-hourly data, which is essential for the planning of RECs, is subordinated to the deployment of smart meters. Many European countries have established a goal to reach a 80 % smart meter penetration rate by 2020 [21], However, this target that has been postponed to 2024 for some of them [22]. According to a recent European study, countries including Austria, Greece and the UK have lagged behind in their large-scale rollout scheduled for 2020 [21]. Moreover, according to a recent update [22], only thirteen European Union countries (Sweden, Denmark, Finland, Estonia, Spain, Norway, Luxembourg, Latvia, Italy, France, Malta, Slovenia and the Netherlands) have achieved the 2020 target. Countries such as Portugal, Austria, the UK and Ireland have set the target of 80 % penetration by 2024, while Belgium, Croatia, Poland, Slovakia, Lithuania and Hungary are at the initial stage of their rollout. Moreover, Bulgaria, Cyprus, the Czech Republic, Germany and Greece have installed very few or no smart meters at all. Consequently, it is evident that the deployment of smart meters in more than half of the countries of the European Union is behind schedule, or has not yet started. This fact reveals the impossibility of obtaining high temporal resolution data in many EU regions, especially considering that even in countries where rollout targets have been recently achieved, it will require time to collect and deploy complete datasets useable for planning processes.

### 1.2. *Literature review on clustering, load profiling and classification techniques*

In this section, we delve into the key methods and algorithms related to load profiling, load pattern recognition, and classification models. We propose a methodology that relies on the integration of these three processes to address the gap identified in Section 1.1, namely the lack of high-resolution data.

Regarding the topic of clustering, many studies have demonstrated the greater effectiveness of k-Means clustering than other methodologies in the context of load pattern recognition using large datasets [22–24]. Indeed, k-Means clustering is often employed alongside evaluation metrics to determine the optimal number of clusters, as proposed by Pérez-Chacon et al. [24]. A different use of k-Means clustering was proposed in the study conducted by Wang et al., which presented a two-step methodology that combined k-Means clustering with Dynamic Time Warping Clustering (DTWC) [25]. The aim of this approach was to address the challenges involved in identifying the typical load profiles of

**Table 1**
Characteristics of the datasets employed in recent research works related to RECs.

| References | Research purposes | REC development phases | Methods | Data type | Data resolution |
|---|---|---|---|---|---|
| [10–12] | REC simulation and optimization | "Design" | MILP / Linear programming | Real industrial, agricultural, tertiary, grid, household, and commercial activity data | Hourly |
| [14,15] | REC simulation, optimization and evaluation of the sharing mechanism | "Design" and "Operation" | Energy and economic-based simulation | Real and synthetic residential and commercial activity data | Hourly / Sub-hourly |
| [13, 16–18] | REC simulation, optimization and management | "Design" and "Operation" | Energy Hub / Time Delay Neural Network / Energy-based simulation | Real residential and building data | Hourly / Sub-hourly |

buildings, and it demonstrated the effectiveness of the two-step clustering process compared to simple k-Means clustering. However, it may only be worth employing such a clustering process in contexts in which a large number of typical load profiles may be present. Similarly, Fang et al., evaluated different clustering algorithms for load pattern recognition [26], including k-Means clustering, Hierarchical clustering, a Gaussian Mixture Model, a C-vine Copula Mixture Model, and an R-vine Mixture Model, in combination with various classifiers such as K-Nearest Neighbors (KNN), Classification and Regression Tree, Naive Bayes Classifier (NB), and Random Forest Classifier, and they demonstrated that a combination of clustering algorithms and a Random Forest Classifier outperforms other combinations, in terms of classification performance.

The research work conducted by Pérez-Ortiz et al., regarding classification techniques, provided a comprehensive review of classification algorithms and their applications in various Renewable Energy (RE) domains [27]. This paper primarily focused on offering an extensive analysis of classification techniques, along with their related learning paradigms, as did Neelamegam et al., who presented a comprehensive summary of the classification algorithms used in the data mining field [28]. These two papers provide in-depth analyses of various classification techniques, and highlight their strengths, limitations, and applicability in different domains. Moreover, Neelamegam et al. also discussed various evaluation measures and validation techniques that could be used to assess the performance of classification models, and emphasized the significance of such metrics as accuracy, precision, and recall.

Regarding the theme of load profiling, Alrawi et al. conducted a study on a data-driven approach to examine users' electricity consumption habits [29]. The main findings of the study demonstrate that the main determinants of electricity consumption are the size of the building, whether the occupants pay electricity bills or not, and the type of air conditioning system. However, it is not easy to obtain information regarding structures, appliances, and users' habits during the development of a Renewable Energy Community project, and for this reason, employing load profiling techniques that rely on such information is not a viable approach within the context of RECs. Similarly, the article written by Piscitelli et al. explored the implementation of a non-intrusive approach to address a customer classification task [30] in the building domain. They developed a classification tool to predict the typical monthly load profiles of new customers. The classification model makes use of a Globally Optimal Decision Tree algorithm, which is trained using predictive attributes derived from the monthly energy bills of each customer, as well as additional information gathered through a phone survey. Furthermore, the developed approach allows the magnitude of typical load profiles to be estimated using energy bill data through specific scaling coefficients. Nevertheless, as previously stated concerning the work of Alrawi et al., the additional information obtained, in this case, through phone surveys, which is related to the occupants' habits and working time, is not easily available in the context of simulating and designing an REC project. Lazzeroni et al. proposed a data-driven approach to predict load profiles for various categories of end users (e.g., for workdays, Saturdays, holidays) using minimal input data, that is, monthly electricity bills [31]. They relied on a K-Nearest Neighbors algorithm to predict typical load profiles on the basis of the similarity between energy bills. They then evaluated the performances of their approach by comparing it with two benchmarks: one based on Standard Load Profiles (SLPs) and another that employed clustering and decision rules to assign new customers to representative load profiles on the basis of their time-of-use energy bill. Their results highlight that the proposed method outperformed the other ones concerning some error metrics and for each end-user category, with just a few exceptions, thus emerging as a valid alternative to load profiling methodologies based on clustering and classification processes.

Considering the Literature Review on clustering, classification, and load profiling techniques, k-Means emerges as the most suitable clustering methodology for big data analysis. Moreover, the classification model coupled with this clustering approach that exhibits the best performance is the Random Forest method. Additionally, the examined load profiling methodologies made use of datasets enriched with additional information, often pertaining to appliances, habits, and occupational activities, except for [31], which proposed a methodology to reconstruct an electrical profile using only monthly electricity bills, utilizing an alternative approach which, however, may be not suitable to reconstruct the electrical load profiles of aggregates in the REC context. In short, the examined research papers proposed models that utilize datasets enriched with additional information, such as the occupants' habits, work activities, and appliances, which is not easily obtainable during the "design" phase of an REC. Furthermore, the analyzed research works pertaining to the context of load profiling were all aimed at reconstructing the load profile of individual users, which is beyond the scope of RECs, where analyses are conducted at an aggregate level.

### 1.3. Research gap, novelty and paper organization

As shown in the previous sections, the recent studies related to REC projects rely on datasets that contain high-resolution data, including additional information regarding the users' habits, work activities, and the appliances. A lack of these data affects the "design" of a new REC project, even in countries where there is a high number of smart meters due to the time it takes to obtain suitable and available data. Additionally, the research studies in the literature that tackled load profiling issues were not specificity aimed at evaluating the key energy parameter of an REC, namely shared energy. For these reasons, given the limited availability of high-resolution data, especially pertaining to the residential sector, the challenge of obtaining additional information, such as the users' work activities and habits, and the lack of existing studies that apply data-driven load profiling techniques to assess key energy parameters of RECs, the aims of this research are to:

- Propose a data-driven methodology, trainable using datasets from different geographical areas, that can be used to reconstruct the hourly electricity consumption of residential users at an aggregate level.
- Reconstruct the hourly electrical consumption of residential users by employing non-intrusive machine learning techniques that rely only on easily obtainable data, such as monthly consumption data, with no additional information related to the users' work activities, habits or appliances.
- Employ the proposed methodology to evaluate the shared energy in a Renewable Energy Community project, in a realistic context of data scarcity, and to propose a model that effectively enables potential aggregators to accurately estimate the shared energy of an REC.

The following sections of this study are organized as follows. Section 2 provides a detailed description of the utilized dataset. Sections 3 and 4 present the description of the proposed methodology and the results obtained from applying the proposed methodology to a specific case study. Section 5 focuses on the discussion and commentary of the achieved results and offers a critical and objective analysis of the outcomes and accomplished objectives. Finally, Section 6 gives a summary of key insights and an outlook of future research.

## 2. Data

The present research work made use of the publicly available Low Carbon London project dataset. The project was conducted by UK Power Networks and developed in London between November 2011 and February 2014 [32]. The project database consists of an extensive time-series dataset of the electrical consumption of 5567 residential users in London for the given period, of which 1122 received an experimental dynamic ToU tariff during 2013, expressed in kWh and measured every 30 min. The consumption measurements can be

considered accurate, with no more than 3.4 % of the measurements missing for any half-hour measurement period, and an average missing measurement per period of just 0.17 % during the 2013 trial year [32]. Within the data set are two groups of customers. The first is a sub-group, of approximately 1100 customers, who were subjected to Dynamic Time of Use (dToU) energy prices throughout the 2013 calendar year period. The remaining sample of approximately 4500 customers energy consumption readings were not subject to the dToU tariff. The database was validated before its publication and cleansed of any missing and/or negative values.

A preliminary data analysis showed that the data collection was inconsistent throughout the project, with some years providing more data than others, as shown in Table 2.

In 2011, data was only collected from 7.9 % of the total users involved during the four years of the project activity, and only two months had at least one electrical consumption measurement per user. The year 2012 was the year with the most comprehensive data collection, with 99.8 % of the total users involved and with twelve months having at least one electrical consumption measurement per user. The year 2013 had slightly fewer users than 2012 and also had twelve months with at least one electrical consumption measurement per user. Finally, 2014 had a high percentage of involved users, but only two months with at least one measured data per user. Furthermore, the month of March appears to be lacking in data primarily for the entire lifespan of the project. Therefore, this month was excluded from the subsequent analyses.

## 3. Modeling framework and methods

### 3.1. Modeling framework

This study introduces a modeling framework used to estimate the hourly electrical consumption of residential user aggregates using monthly consumption data through non-intrusive machine-learning techniques. The aim of the proposed framework is to use the simulated hourly electrical consumption to evaluate the shared energy in a REC project, enabling potential aggregators to accurately estimate the shared energy in a realistic context of data scarcity, allowing them to conduct pre-feasibility studies and economic estimates in a more robust and reliable way. The proposed methodology can be divided into two major processes: model creation and case study application. The model creation phase involves creating a model that can reconstruct the hourly electrical profile at the aggregate level. This model receives real or synthetic data, with at least an hourly resolution, as input, which are required to train the model. The case study application phase involves reconstructing the hourly profile of the desired aggregate electrical load, and then calculating the primary evaluation metrics for the obtained load profile. This phase requires real or synthetic monthly electrical consumption data pertaining to the users for whom the aggregate electrical profile has to be reconstructed, as input. The model creation phase generally consists of five steps: *i) Pre-processing, ii) Data filtering. iii) Load Pattern Recognition, iv) Classification, and v) Scaling Coefficient Evaluation,* as depicted in Fig. 2.

The first step consists in applying an extensive pre-processing to the available raw data. This process is based on the methodology proposed in [32] and is aimed at eliminating elements of the dataset that could negatively impact the performance of the subsequent models and processes. This phase is primarily divided into two sections. The first section, known as data characterization, is aimed at enhancing the understanding of the dataset through appropriate visualization techniques; the second section, namely data preparation, focuses on cleansing the dataset from missing values, outliers, and inconsistencies. A data filtering process is then undertaken with the purpose of further refining the dataset by removing infrequent load profiles. This is achieved through a segmentation of the dataset into two steps: k-Means clustering followed by an analysis of the populations of the clusters to eliminate the least populated ones. The load pattern recognition process is then undertaken. The aim of this process is to identify the typical electrical load shapes of the residential users, using k-Means clustering. In a similar way to the data filtering process, an analysis is conducted on the population of the obtained clusters with the aim of eliminating any load profiles that are not representative of the users' consumptions. Next, the classification process is conducted to create a model that is able to assign one of the previously identified typical load shapes to a generic user for each month of the year, basing this assignment on monthly electrical consumption. In other words, the process involves implementing a Random Forest method, followed by an iterative optimization aimed at identifying the hyperparameters of the model that maximize performance. Finally, a scaling coefficient evaluation process is undertaken to evaluate the scaling coefficients that should be used to transform the typical load profiles attributed to users for each month of the year into electrical load profiles. This is a crucial set and requires the definition of scaling coefficients for each user and for each month of the year. At the end of the process, the obtained scaling coefficients are evaluated to reconstruct the electrical load profiles of each user on a monthly basis. Therefore, assuming that, for a single user, the daily electricity consumption corresponds to the obtained monthly load profiles for each month, the hourly electricity consumption for the entire year can be reconstructed. A clear and concise explanation of the proposed methodology and the obtained results is provided in the next sections, and detailed descriptions of the models and the underlying assumptions are presented.

**Table 2**
Preliminary dataset analysis.

| Parameter | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|
| Number of involved users | 354 | 4433 | 4411 | 4065 |
| Relevant Months[1] | 2 | 12 | 12 | 2 |
| Percentage of involved users[2] | 7.9 % | 99.8 % | 99.3 % | 91.5 % |

[1] Months with at least one consumption data detected for each user.
[2] Percentage of users to the total number of users involved in the project.

### 3.2. Pre-processing

Data pre-processing is an essential step in all data-driven model implementations, as it is mandatory to ensure accurate results and reliable analyses, and it includes the preparation and transformation of data into a suitable form for the mining procedure. The aim of data preprocessing is to reduce the size of the data, to find the relationships between the data, to normalize the data, to remove any outliers, and to extract features. It involves several techniques, such as data cleaning, transformation, and reduction. The process can be divided into two main tasks, as shown in Fig. 3: data preparation and data characterization. The data characterization phase is necessary to obtain an initial comprehension of the data, particularly considering the difficulty of visually representing multidimensional data in a comprehensive manner. The data preparation phase is an important step which allows to obtain clean, well-structured and suitable data.

Before embarking on the processes of data preparation and data characterization, a preliminary data selection is performed to identify the subsection of the dataset that corresponds to the most complete year, in terms of the number of involved users and months with complete measurements. The dataset subsection corresponding to 2012 was chosen, as it emerged as the most populated subsection of the dataset. In addition, only the data from users with complete electricity consumption measurements, i.e., for the entire year, were considered from the 2012 subsection dataset, which led to the selection of 272 users.

In the data characterization phase, an analysis of the probability density function is performed on the hourly consumption measurements, and a frequency analysis is conducted on the total monthly consumption, as shown in Fig. 4, to attain a more comprehensive

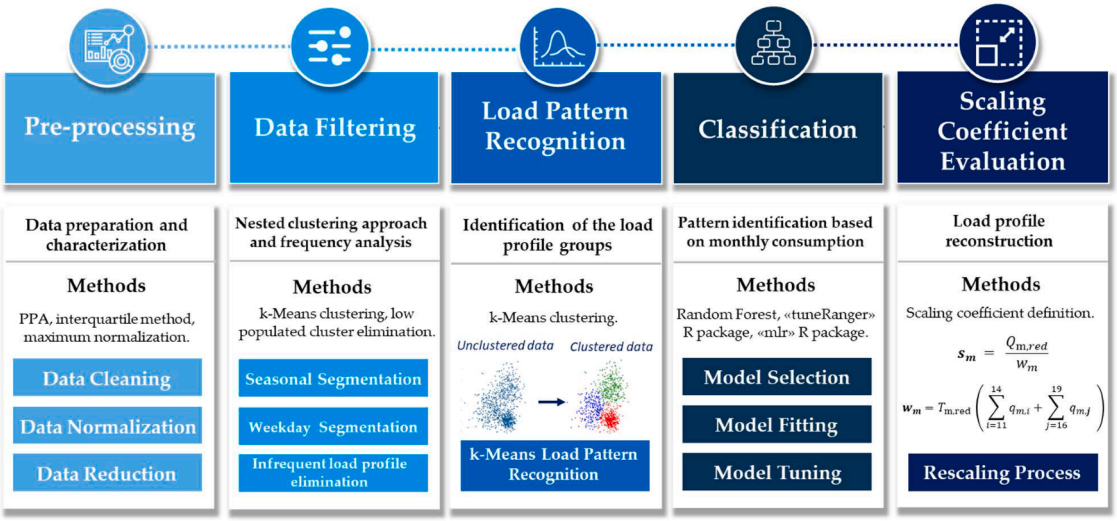



**Fig. 2.** Framework of the model creation phase.

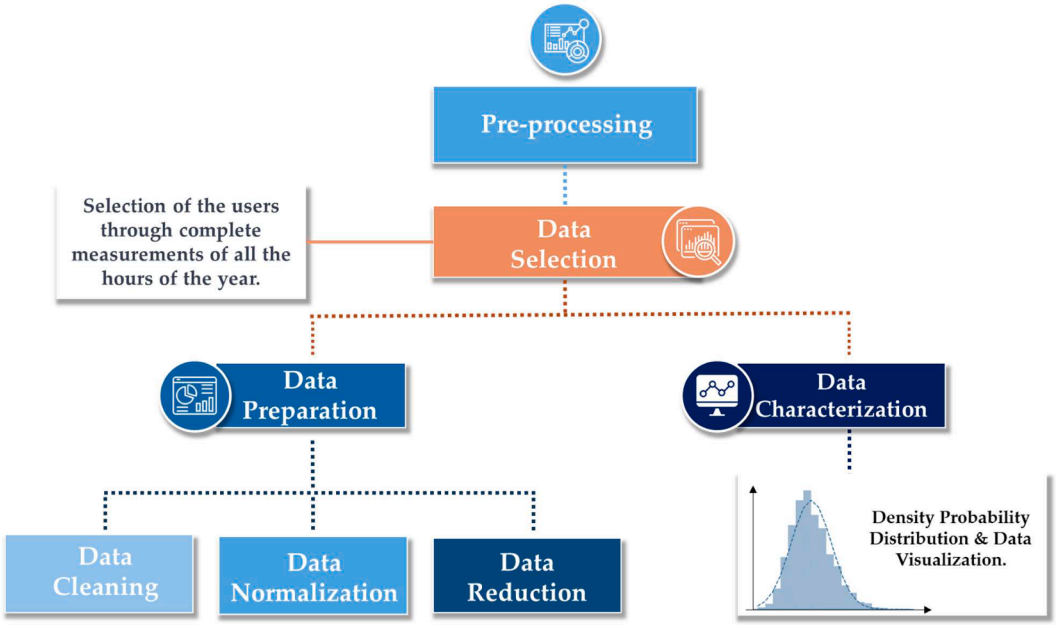



**Fig. 3.** Pre-processing workflow.

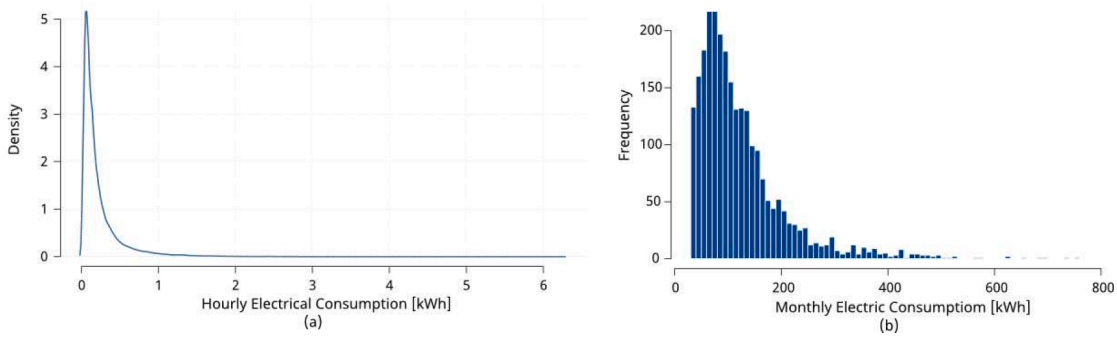



**Fig. 4.** (a) Density probability of the pre-processed data; (b) Monthly consumption frequency analysis of the pre-processed data.

understanding of the dataset.

As illustrated in the figure, the hourly electrical consumption measurements of the pre-processed data primarily correspond to low consumption levels. This can be attributed to two main factors. Firstly, the presence of nighttime electrical consumption values, which are significantly lower than the daytime ones. Secondly, the presence of users with relatively low electrical consumption, as illustrated in Fig. 4b, where most values represent medium-low monthly consumptions. Fig. 4b also highlights some notably high monthly electrical consumption values, which reach close to 800 kWh in some instances. This demonstrates the presence of energy-intensive users during certain months of the year, even though they are in a smaller number, thus underlining the diversity of the users' characteristics within the pre-processed data. The results of the data preparation process, which concludes the pre-processing phase, show that no missing values or outliers were detected. This confirms the information in Section 2 related to the data cleaning and validation processes carried out before the considered dataset was made available to the public.

Subsequently, the data preparation process, which is mainly focused on two issues, is carried out: firstly, the data should be organized into an appropriate form for the subsequent data mining algorithms, and secondly, the data should be processed to attain the best performance and quality of the models involved in the subsequent data mining operations. The data cleaning process involves managing the missing values, inconsistencies, and outliers. Any missing values are handled using a linear interpolation method, which consists of substituting missing values with those on the straight line between two valid measures [33]. Whenever such measures are somewhat distant from each other, linear interpolation is ineffective in handling any missing values, and for this reason the maximum distance between the two valid measures is set equal to 3 during data cleaning process. Outliers are handled on an hourly basis using the "interquartile method". This is one of the simplest and most effective methodologies used for data mining and it requires the calculation of three fundamental parameters [34]:

- The median (Q2), which is the central value of the sampled data distribution.
- Q1, which is defined as the first quartile, or the value below which 25 % of the initial data is present.
- Q3, which is defined as the third quartile, or the value below which 75 % of the initial data is present.

The distance between Q3 and Q1 is defined as the inter-quartile range (IQR), and a new interval, known as the decision range, is defined using this range, with any elements outside of it being identified as outliers.

The Lower and Upper Bounds are defined by Eqs. (1) and (2).

$$Lower\ Bound = Q_1 - 1.5 * IQR \tag{1}$$

$$Upper\ Bound = Q_3 + 1.5 * IQR \tag{2}$$

The data transformation process is then carried out by normalizing the raw data, which is a mandatory step for certain machine learning techniques [30]. In other words, the data are normalized using a maximum scaling approach, as shown in Eq. (3).

$$\widehat{x}_{i,m} = \frac{x_{i,m}}{\max\left(x_{i,m}\right)} \tag{3}$$

where:

- $x_{i,m}$ represents the monthly load profile of the i-*th* user for the m-*th* month;
- $\max(x_{i,m})$ is the maximum value of the load profile of the i-*th* user for the m-*th* month.

### 3.3. Data filtering

In this section, a specific process is employed to eliminate any elements within the pre-processed data that could further adversely affect the performance of the implemented data-driven and machine learning models. The process is divided into two phases that are conducted to identify any infrequent daily load profiles, in terms of intensity and shape. In the first phase, the daily profiles pertaining to months characterized by a low electricity consumption are eliminated for each user present in the pre-processed data using a threshold value ($TV_{months}$) defined according to Eq. (4), which represents an approximation of the monthly electricity consumption of residential users in England.

$$TV_{months} = 0.10 * \frac{Q_{UK,avg}}{12} \tag{4}$$

where:

- $Q_{UK,avg}$ is the average yearly electric consumption of residential users in England, which is equal to 3800 kWh for the year 2014 [35].

Thereafter, any infrequent daily load profiles, in terms of load shape, are identified using a nested approach, which involves first segmenting the pre-processed data by season, as summarized in Table 3, and then by days of the week, followed by k-Means clustering, which is performed on each of the identified subsections, as depicted in Fig. 5.

This approach enables any subgroups within each subsection that contain infrequent load profiles or profiles associated with a restricted group of users to be identified using the threshold values defined by Eqs. (5) and (6).

$$TV_u = 0.05 * N_{max,u} \tag{5}$$

$$TV_p = 0.05 * N_{max,p} \tag{6}$$

where:

- $TV_u$ represent the threshold value related to the number of users within each clusters.
- $TV_u$ represent the threshold value related to the number of profiles within each clusters.
- $N_{max,u}$ is equal to the maximum number of users within each subsection.
- $N_{max,p}$ is equal to the maximum number of profiles within each subsection.

The least populated clusters are identified on the basis of the threshold values calculated for each subsection of the dataset, as depicted in Fig. 6, and the daily load profiles within each cluster are subsequently excluded from further analysis.

The number of clusters used during the clustering process performed for each subsection is identified through the Davies–Bouldin Index (DBI). The DBI [36] is a cluster validity metric which is based on the concept that, in order to obtain a good partition, the inter cluster separation as well as the intra cluster cohesion should be as high as possible. The DBI is evaluated according to Eq. (7).

$$DBI = \frac{1}{K}\sum_{k=1}^{K} \underset{k\neq 1}{max}\left(\frac{\delta_k - \delta_i}{d_{k,i}}\right) \tag{7}$$

**Table 3**
Seasonal dataset segmentation.

| Season | Months |
|---|---|
| Summer | June, July, August |
| Winter | December, January, February |
| Spring | March, April, May |
| Autumn | September, October, November |

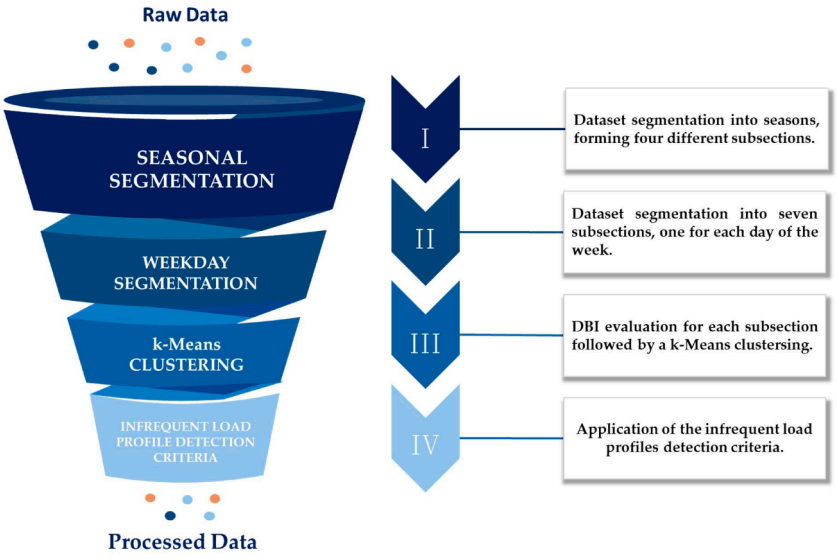



**Fig. 5.** Workflow of the infrequent daily load profile detection.

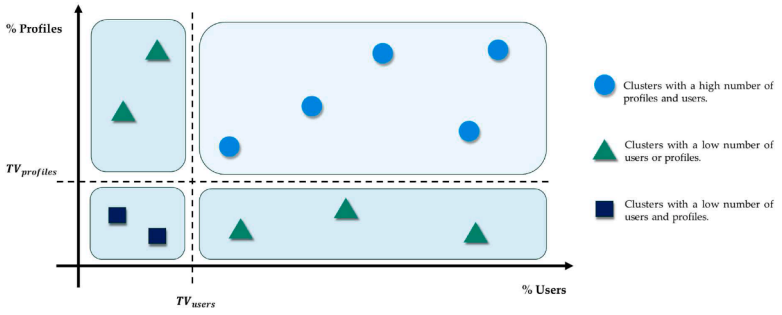



**Fig. 6.** Infrequent load profile detection criteria.

where:

- $k$ is the number of clusters.
- $d_{k,i}$ is the Euclidean distance between the centroids of clusters $C_k$and $C_i$.
- $\delta_k$, $\delta_i$ are the standard deviations of the distances between objects in clusters $C_k$ and $C_i$.

The DBI basically measures the similarity between clusters considering their internal distribution and degree of separation. A low index value indicates a good division, while a high index value means that the division is not optimal, the generated clusters are not easily distinguishable, and there are significantly different elements within clusters.

An example of applying the criterion used to identify infrequent daily load profiles is presented in Fig. 7, where infrequent daily profiles can be identified for Sundays in the autumn season and Thursdays in the spring season.

The optimal number of clusters in Fig. 7a, identified through the DBI index, is 16, while the threshold values for the number of users ($TV_u$) and the number of profiles ($TV_p$) are 8 and 120, respectively. In this case, clusters 6, 7, 8, and 9 have several profiles below $TV_p$ and were therefore excluded from the pre-processed data, while there were no cluster with a

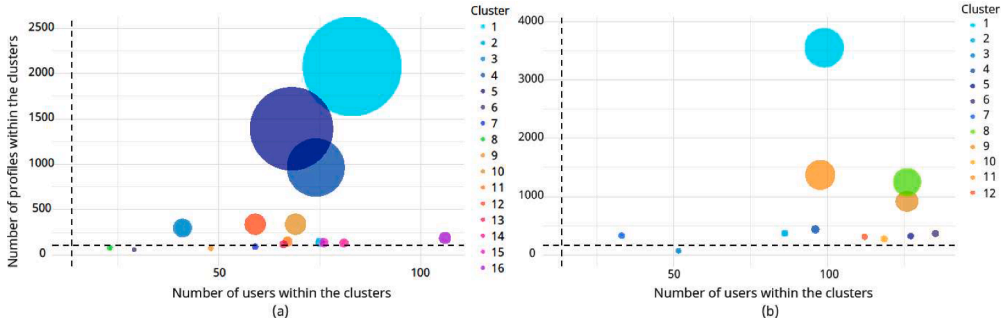



**Fig. 7.** (a) Detection of the daily-scale outliers for Sundays in the autumn months. Clusters within the threshold values were considered infrequent or non-representative and excluded from the analysis; (b) Detection of the daily-scale outliers for Thursdays in the spring months. Clusters within the threshold values were considered infrequent or non-representative and excluded from the analysis. The size of the circles in both figures represents the number of profiles and users within a cluster.

number of users below $TV_u$. Similarly, the optimal number of clusters in Fig. 7b, obtained through the DBI, is 12, while the threshold values for the number of users and the number of profiles are 7 and 220, respectively. The only cluster below the threshold values in this subset of the dataset is cluster 3, and its profiles were excluded from the pre-processed data. As can be observed, the number of clusters below the threshold values is influenced to a great extent by the number of clusters chosen during the clustering process. In this regard, the DBI helps determine a suitable number of clusters that are neither too small, which could lead to incorrect profile distinctions, nor too large, which could result in an excessive division of the dataset subsection. Finally, as can be seen in the two aforementioned examples, the most impactful variable when identifying anomalous daily profiles is $TV_p$. Indeed, the latter accounts for the elimination of 87 % of the clusters throughout the entire process, while the remaining portion is determined by the threshold value related to the number of users and the combination of the two threshold values.

At the end of the data filtering process 7223 daily load profiles were identified as infrequent, as shown in Table 4, which presents the number of identified infrequent daily load profiles for each season and their respective relative percentages.

As can be seen, the highest number of identified infrequent daily load profiles, which is 2850 load profiles, occurs during the summer season. Meanwhile, the smallest number of infrequent daily load profiles, which is 750 load profiles, is observed during the winter season. This could be attributed to the fact that users tend to exhibit more variable electrical load patterns during the summer due to the different use of appliances and space cooling systems, which may be less pronounced in the winter season. However, the analysis of the relationship between the electrical consumption changes and seasons is beyond the scope of this research work.

In summary, the data filtering process resulted in an overall elimination of 12.8 % of the dataset, with 8.0 % identified as infrequent daily load profiles and 4.8 % categorized as low monthly consumption data, as shown in Table 5.

### 3.4. Load pattern recognition

After the pre-processing and data filtering processes, a k-Means clustering is performed on the monthly electrical consumption. Once again, the optimal number of clusters is evaluated using the DBI. Similar to the data filtering process, clusters containing a low number of profiles or users are eliminated using the same logic. In this case, the used threshold values are set to 10 % of the number of profiles and users. However, this time, unlike the clustering performed during data filtering, the Hartigan and Wong algorithm is utilized. The Hartigan and Wong algorithm is computationally more expensive than the previously used Lloyd algorithm, but it can provide a more stable and robust clustering, especially when the number of clusters is high, and the initial centroids are chosen non-optimally [37,38].

In this phase, the optimal number of clusters defined through DBI was chosen as 5, whose variation in relation to the number of clusters (from 4 to 20 clusters) is illustrated in Fig. 8, and the obtained typological load patterns identified through k-Means clustering are depicted in Fig. 9, in which five different load patterns can be observed.

**Table 4**
Daily-scale outliers identified for each season.

| Season | Infrequent daily load profiles | Profile percentages[1] |
|---|---|---|
| Summer | 2850 | 39.5 % |
| Winter | 750 | 10.4 % |
| Spring | 1981 | 27.4 % |
| Autumn | 1642 | 22.7 % |

[1] Percentages related to the maximum numbers of daily load profiles for each season.

**Table 5**
Results obtained from the data filtering process.

| Dataset subsection | Percentages |
|---|---|
| Valid Data | 87.2 % |
| Excluded Data | 12.8 % |
| Infrequent Daily Load Profile Data | 8.0 % |
| Low Monthly Consumption Data | 4.8 % |

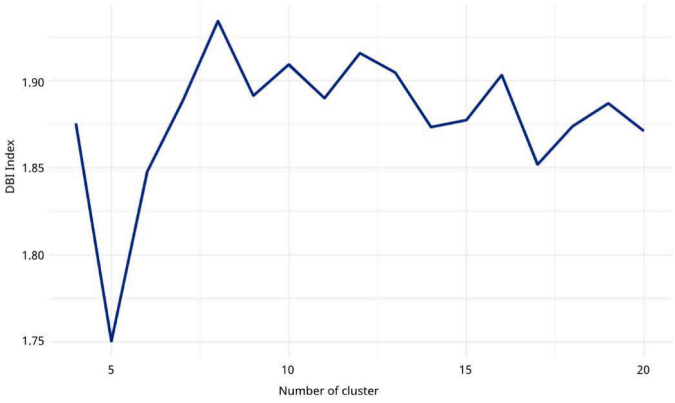



**Fig. 8.** Evaluation of the optimal number of clusters using the DBI index for the load pattern recognition process.

The first obtained typological profile is characterized by a rapid decline from 00:00 until approximately 05:00, followed by an upward trend, while a consistent load is maintained in the middle of the day. Around 15:00, there is a sudden surge in load, which peaks at around 19:00, before descending again. The second typological load profile is distinguished by a rapid increase in load, starting at around 05:30, which reaches a first peak at 09:00. In the subsequent hours, the load is maintained at an approximately constant value until around 14:30, where a slight descent occurs, followed by a rapid increase that leads to a second peak at 18:00. The load then decreases significantly during the nighttime hours. The third typological load profile is significantly different from the others. It exhibits a much lower average load value than the others and features two closely spaced peaks during the nighttime hours, followed by a rapid decline at around 07:30, after which it maintains a constant load significantly lower than the nighttime hours. Instead, the fourth typological load profile has a generally higher average load value than the others. It is characterized by a slight decline during the nighttime hours, followed by a slight increase in the load that remains constant during the midday hours, reaching a slight peak at around 19:30 and then gradually decreasing. Finally, the fifth load profile is similar to the first one, but some timeframes differ. In this load profile, the decline in load during the nighttime hours is less pronounced, and the load during the midday hours, which remains relatively constant, is lower than the first typical load profile. Furthermore, the localized peak during the nighttime hours occurs slightly earlier, at around 18:30.

After identifying the typical load patterns, as previously stated, a further check was performed to exclude clusters with a low number of profiles and the users within these clusters, adopting the same procedure used to eliminate the infrequent daily load profiles during the data filtering phase. In short, the threshold values related to the number of users and profiles within the clusters were evaluated as previously defined by Eqs. (5) and (6), setting the threshold values as 10 % of the number of profiles and users and used to identify clusters containing low populated clusters, as illustrated in Fig. 10.

In this case, the threshold values pertaining to the number of users and the number of profiles were set to 60 and 262, respectively. As is evident from the graph, the least populated cluster is the number 3,

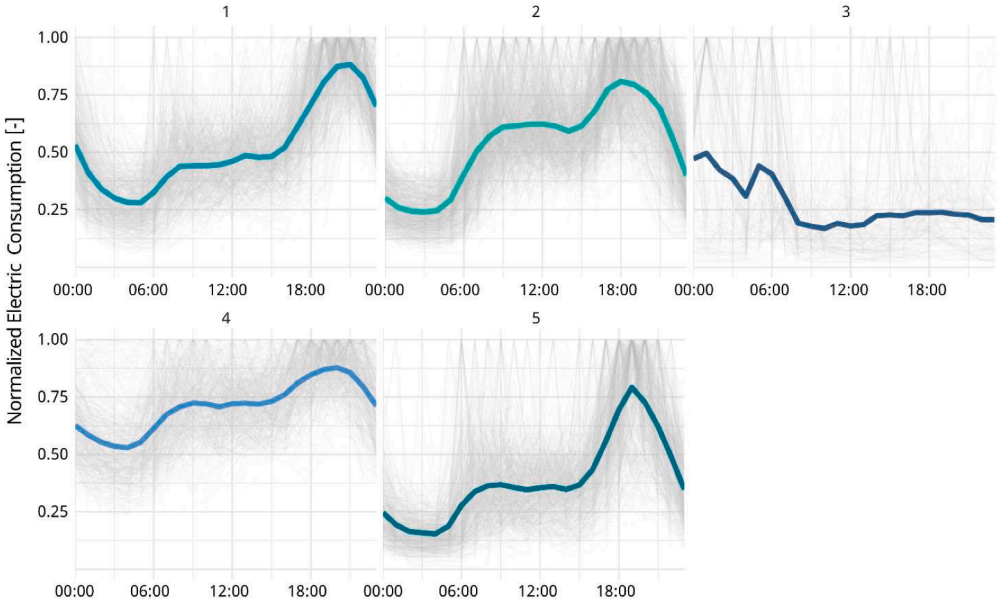



**Fig. 9.** Typical electrical load patterns obtained through the load pattern recognition process.

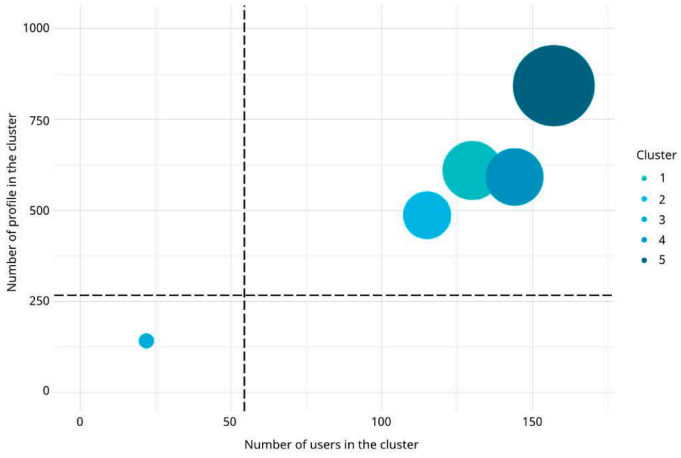



**Fig. 10.** Elimination of the low populated clusters for the load pattern recognition process.

which corresponds to the third typological load profile, and it is significantly different from the others, as previously mentioned.

To summarize, typological load profiles are obtained as centroids of the four remaining clusters at the end of the load pattern recognition process. Those patterns are subsequently used to train a supervised classification model to assign a consumption pattern to users based on their monthly consumption values.

### 3.5. Classification

The main objective of the classification process is to create a model that is able to identify, using monthly consumption values, the electrical load pattern on a month-by-month basis throughout the entire year, as shown in Fig. 11. The ToU is evaluated considering the London Power (LOND) company's ToU, as outlined in Table 6. These consumption bands are considered the same for each day of the year.

The creation of the classification model involves the following steps: *i) selection of the classification model; ii) definition of the training variables; iii) definition of the training and testing datasets; iv) setting up the model; v) tuning the model.* The selection of the classification model is one of the most important phases of the process. According to the study conducted by Fang et al., the model that best fits k-Means clustering for classification processes is the Random Forest model [26]. Therefore, it was decided to implement a Random Forest model using the "RandomForest" package in the R development environment. This category of classifiers belongs to the realm of ensemble-based machine learning models, which employ a collection of decision trees (DTs). DTs are employed to classify patterns by utilizing a series of precisely defined rules. They include a classification algorithm that resembles a tree structure, where attributes of the dataset are depicted as internal nodes, decision rules are depicted as branches, and the outcomes of the combined decisions are depicted as leaf nodes. Several decision trees are trained in Random Forest models by randomly choosing subsets of the dataset that have the same size as the original training set. A final class of the test object is then determined by combining the predictions from the various decision trees. The typology of Random Forest that has here been used is called "Ranger Forest", which is a fast implementation of the Random Forest model that is particularly suitable for high dimensional data. After selecting the model, the training variables are defined. As previously stated, the main objective of the classification process it to assign a consumption pattern on the basis of the monthly consumption data. For this reason, the variables used to train the model are the monthly consumption values, the ratio between the monthly consumption values within the ToU time frames, and the total monthly consumption, as shown in Table 7.

It was decided to incorporate some other variables during the training phase to enhance the accuracy of the model. The new variables employed are the same as those shown in Table 7, albeit corrected by a correction factor, according to Eq. (8), except for the electrical consumption in the "Red", "Amber", "Green" ToU rates and the total monthly electrical consumption.

$$f_{m,t/j} = \frac{T_{m,t}}{T_{m,j}} \tag{8}$$

where:

- $T_{m,t}$ represent the number of hours of the t-*th* ToU rate over the m-*th* month.
- $T_{m,j}$ represent the number of hours of the j-*th* ToU rate over the m-*th* month.

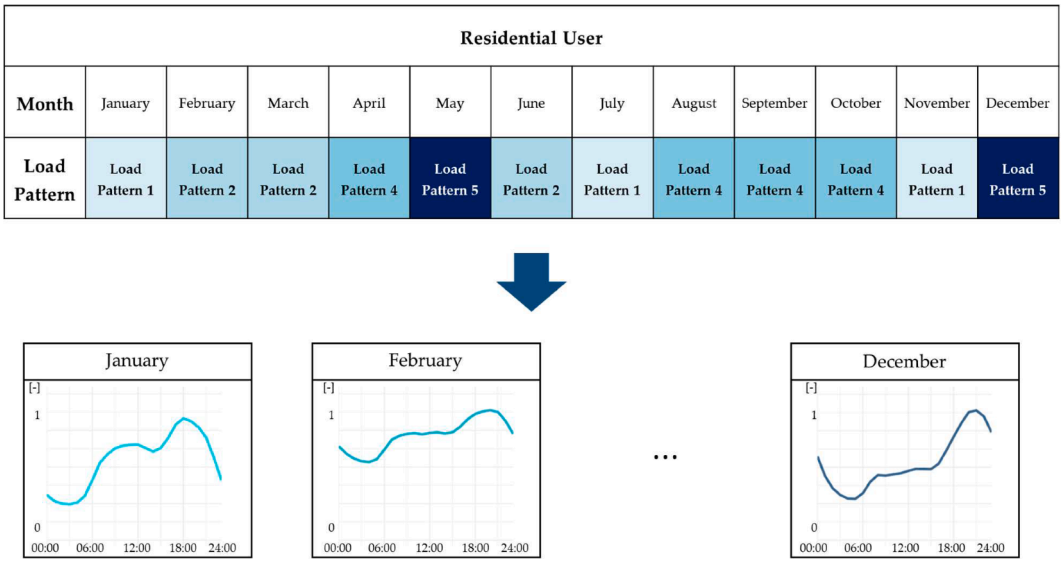



**Fig. 11.** Classification task carried out for a generic residential user.

**Table 6**
The London Power company's ToU.

| Time of use | Hours |
|---|---|
| Red | 11:00–14:00 |
| | 16:00–19:00 |
| Amber | 07:00–11:00 |
| | 14:00–16:00 |
| | 19:00–23:00 |
| Green | 00:00–07:00 |
| | 23:00–24:00 |

**Table 7**
Training variables used to train the classification model.

| Training variable | Description | Unit |
|---|---|---|
| Amber/Total | Electrical consumption of the "Amber" ToU timeframe on the monthly electrical consumption | [-] |
| Red/Total | Electrical consumption of the "Red" ToU timeframe on the monthly electrical consumption | [-] |
| Green/Total | Electrical consumption of the "Green" ToU timeframe on the monthly electrical consumption | [-] |
| Amber/Red | Electrical consumption of the "Amber" ToU timeframe on the electrical consumption of the "Red" ToU timeframe | [-] |
| Green/Amber | Electrical consumption of the "Green" ToU timeframe on the electrical consumption of the "Amber" ToU timeframe | [-] |
| Green/Red | Electrical consumption of the "Green" ToU timeframe on the electrical consumption of the "Red" ToU timeframe | [-] |
| Green | Electrical consumption of the "Green" ToU timeframe | [kWh] |
| Amber | Electrical consumption of the "Amber" ToU timeframe | [kWh] |
| Red | Electrical consumption of the "Red" ToU timeframe | [kWh] |
| Total | Electrical monthly consumption | [kWh] |

The correction factor, *f,* takes into account the ratio between the temporal extension of the ToU rates over a month, namely the ratio between the number of hours of two different ToU rates, which is information that is usually hard to transfer to the model using only the monthly consumption values.

Subsequently, the training and testing datasets are defined. The characteristics of these datasets are summarized in Table 8.

After having identified the model, and after having defined the training variables, and the training and testing datasets, it is necessary to determine the hyperparameters that define the best model, in terms of classification performance. Random Forest classification models are characterized by three hyperparameters: the number of globally generated trees, the number of variables used for each split, and the minimum elements in the leaf nodes.

**Table 8**
Training and testing the characteristics of the datasets.

| Dataset name | Percentage of the overall dataset | Description |
|---|---|---|
| Training Dataset | 70 % | Seventy percent of the overall dataset randomly selected to train the classification model. |
| Testing Dataset | 30 % | Thirty percent of the overall dataset randomly selected to test the performance of the classification model. |

The number of globally generated trees is one of the most important parameters. It is closely related to the size of the database and the trade-off between accuracy and computational cost. In general, the number of trees can be chosen arbitrarily, although, in practice, a number of trees between a few hundred and a few thousand can provide good results for most problems. Furthermore, the accuracy of the model tends to improve as the number of trees increases, although, after a certain point, increasing the number of trees does not lead to any significant improvement in accuracy. The number of variables used for each split indicates the number of training variables that are used to create a split rule at each step. Using a different number of split variables may lead to a significantly different performance of the model. Finally, the minimum number of elements in the leaf nodes represents the minimum limit value of the elements within a terminal node. This value is closely related to the size of the generated tree. Large trees that contain many split nodes are obtained for a low minimum number of elements in the leaf nodes. Conversely, smaller trees are obtained with a limited number of split nodes for a high number of these elements in the leaf nodes. In our case, "tuneRanger" and "mlf" packages were used in the R programming environment to find the best combination, in terms of classification performances and hyperparameters, and 200 iterations were performed in which the different combinations of hyperparameters were evaluated, with a fixed number of generated trees that was set equal to 4000.

The other parameters that are used to define the model and the iterative optimization process are: *importance, split rule,* and *measure*. The *importance* parameter identifies the metric used to evaluate the importance of each training variable and is set in order to reduce the

measure of impurity, defined by the Gini Index, within groups of observations. The *split rule* parameter defines the rule used to select the best variables, in terms of classification performances, at a split node, and the Gini Index was used for this specific model. The *measure* parameter defines the KPIs considered to identify the models with the best classification performances during the iterative hyperparameter optimization process. In this case, the accuracy of the model was defined by choosing the best KPI to evaluate the classification performances.

In synthesis, an iterative optimization process was conducted to obtain the model with the best accuracy to identify the best combination of hyperparameters, using a fixed number of trees equal to 4000, whose results are summarized in Table 9.

The number of variables required to define the rules used by the best Random Forest model was set to 2, while the minimum number of elements in the leaf nodes, which is strictly related to the trade-off between computational costs and model performance, was defined to be 16, which corresponds to less than 1 % of the total number of elements analyzed by the overall model. The performance of the classifier, which refers to its ability to correctly classify in relation to the total number of classifications, was evaluated, and accuracies of 66 % and 63 % were obtained during the training and testing phases, respectively.

The classifier was not much accurate, and this is primarily because the model was trained with variables that were not fully descriptive of the identified load profile types during the load pattern recognition process. As a result, the model struggles to assign the correct load type based on the trained variables only. Additional information would be required, such as the number of occupants, their habits, or even their occupational activities, to improve the performance of the model. However, such information was not considered the present research work, which has been conducted in a context of data scarcity.

A ranking of the predictive variables deemed the most important by the model is presented in Fig. 12, where the importance of the variables was calculated through the Gini Index.

Fig. 12 shows the considered variables in order of importance from top to bottom. The top three most important variable are the ones weighted by correction coefficient “f”. This result shows the effectiveness of providing such additional information to support the Random Forest classification. Furthermore, the figure shows that the absolute values of electricity consumption in different ToU rates and the monthly consumption are associated with a significantly lower level of importance compared to the remaining variables. This implies that, in order to identify the load profile shapes of users, the ratio of electricity consumption values in different ToU rates are much more important than the absolute consumption values.

### 3.6. Scaling coefficient evaluation and case study application phase

This section outlines the key steps involved in calculating the scaling coefficients that are necessary to pass from the typical normalized load profiles, defined in the load pattern recognition phase and assigned to each user for each month during the classification process to real load profiles. This step is the final stage of the model creation phase and is directly interfaced with the case study application phase, as depicted in Fig. 13.

**Table 9**
Hyperparameters of the random forest classifier obtained through an iterative optimization process.

| Hyperparameter | Value | Description |
|---|---|---|
| Number of variables used for each split | 2 | Number of elements necessary to create a rule |
| Minimum elements in the leaf nodes | 16 | Number of elements within a terminal node |
| **Additional Random Forest parameters** | | |
| Importance | Gini Index | |
| Split rule | Gini Index | |
| Measure | Accuracy | |

The coefficients used in the scaling coefficients evaluation process to weigh the magnitude of the monthly consumption for the "Red" ToU rate are calculated, using the values of the typical normalized load profiles obtained during the load pattern recognition process, and are assigned to each user on a month-to-month basis during the classification process. The coefficients used to weight the magnitude of the monthly consumptions for the "Red" ToU rate are calculated according to Eq. (9).

$$w_m = T_{m,red}\left(\sum_{i=11}^{14} q_{m,\,i} + \sum_{j=16}^{19} q_{m,j}\right) \tag{9}$$

where:

- $T_{m,red}$ represents the number of hours of "Red" ToU rate for the m-*th* month.
- $q_{m,i}$ is the value of the normalized typical load profiles at the i *th* hour during the time frame defined by the "Red" ToU, i.e., between 11:00 and 14:00, and between 16:00 and 19:00, for the m-*th* month, which is used to weight the magnitude of the monthly consumption value.

As shown in Eq. (9) and Fig. 14, the values residing within the "Red" ToU time frame were used for each assigned typological load pattern to define the weight coefficients.

A further comparison was made of the four typological load profiles obtained downstream of the load pattern recognition process in Fig. 14a, where their differences, which determine the varying coefficients used to pass from typical load profiles to simulated load profiles, can be appreciated, while graphical representations of the sections of the typological patterns used to calculate the scaling coefficients employed to weight the monthly consumption values are provided in Fig. 14b–e. As can be observed from the graph, Load Pattern 4 emerges as the most distinct typological profile, as it exhibits the highest load intensity. The remaining patterns are similar to each other, particularly Load Pattern 1 and Load Pattern 5, which primarily differ in shape during the nighttime hours and in the temporal location of the evening peak, which is slightly shifted in Load Pattern 1 compared to Load Pattern 5. Additionally, Load Pattern 2 closely resembles Load Pattern 5 during the nighttime hours, but significantly deviates from it during the midday hours, where it displays a higher load intensity.

After obtaining the weight coefficients, the scaling coefficients necessary to pass from the typical normalized load patterns to real load profiles are evaluated using Eq. (10).

$$s_m = \frac{Q_{m,\,red}}{w_m} \tag{10}$$

where:

- $Q_{m,red}$ represents the monthly consumption within "Red" ToU for the m-*th* month.

The reference value for the scaling factor was chosen as the consumption value in the "Red" ToU time frame, as it generally involves periods of higher consumption for residential users, and covers most of the consumption activity time frame, which corresponds to approximately 50 % of the hours during which residential users tend to have medium-high consumption, i.e., between 8:00 and 20:00.The different shapes of the typological load profiles led to different scaling coefficients, as shown in Table 10.

The highest weight coefficients are associated with typological load patterns, namely Load Patterns 2 and 4, that exhibit a greater load intensity than the others. Consequently, their influence will be more significant during the scaling process, while it will be less significant for those months assigned Load Patterns 1 and 5 by the classifier.

At the end of the process, simulated electrical consumption profiles

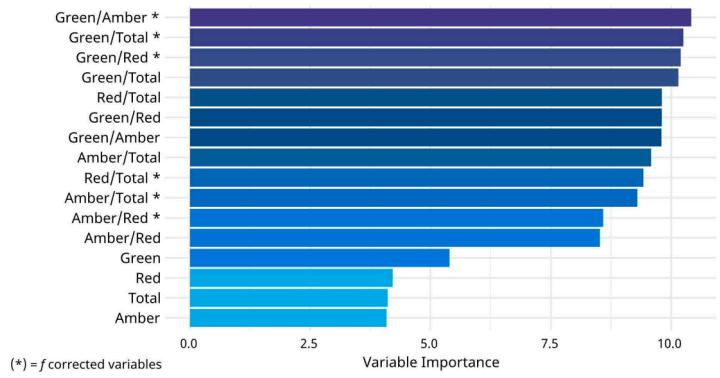



**Fig. 12.** Importance of the variables for the classification process. (*) variables corrected with correction factor "*f*".

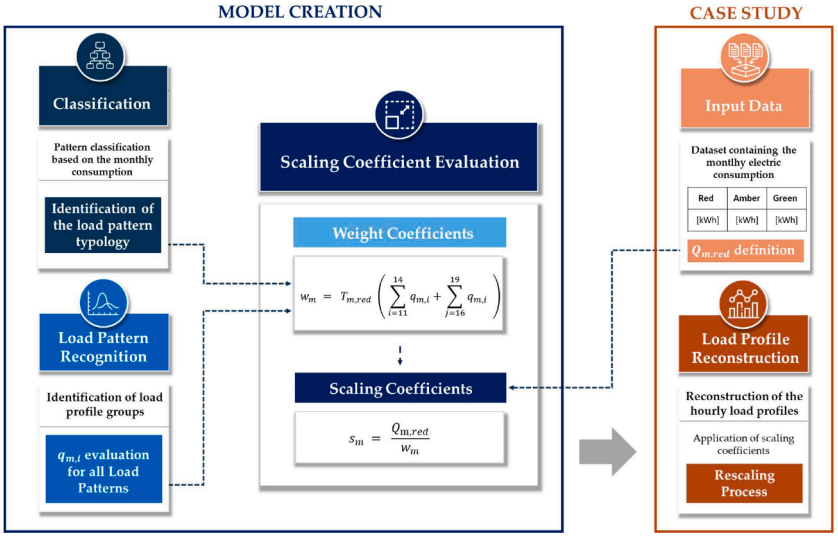



**Fig. 13.** Workflow of the evaluation of the scaling coefficients.

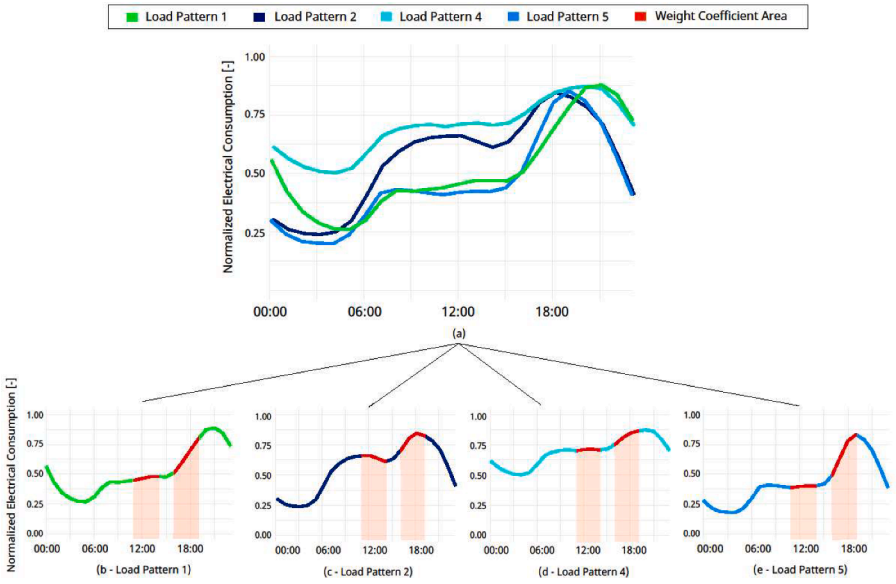



**Fig. 14.** (a) Comparison of typical load patterns; (b) typical load profile values residing within the "Red" ToU timeframe for load pattern 1; (c) typical load profile values residing within the "Red" ToU timeframe for load pattern 2; (d) typical load profile values residing within the "Red" ToU energy consumption range for load pattern 4; (e) typical load profile values residing within the "Red" ToU timeframe for load pattern 5.

**Table 10**
The values considered to calculate the scaling coefficients used to weight the monthly consumption values.

| Load pattern | 11:00 | 12:00 | 13:00 | 16:00 | 17:00 | 18:00 | Weight coefficient ($s_m$) |
|---|---|---|---|---|---|---|---|
| 1 | 0.43 | 0.44 | 0.45 | 0.49 | 0.58 | 0.69 | 18.48 |
| 2 | 0.63 | 0.62 | 0.59 | 0.69 | 0.76 | 0.82 | 24.66 |
| 4 | 0.71 | 0.71 | 0.72 | 0.76 | 0.81 | 0.84 | 27.30 |
| 5 | 0.36 | 0.37 | 0.36 | 0.45 | 0.58 | 0.73 | 11.04 |

are obtained for each month, by multiplying the $s_m$ factor by the previously obtained normalized load profiles, according to Eq. (11).

$$\widehat{Q}_{m,h} = q_{m,h} * s_m \forall i \in A \tag{11}$$

where:

- $A$ = {January, February, …., December};
- $\widehat{Q}_{m,h}$ represents the simulated consumption value, for the m-*th* month, at hour *h*;
- $q_{m,h}$ is the value of the normalized consumption curve, for the m-*th* month, at hour *h*;
- $s_m$ is the scaling coefficient for the m-*th* month.

As previously stated, at the end of this phase, a month-by-month electrical load profile is obtained for each user, as depicted in Fig. 15.

At this stage, average electric load profiles are reconstructed for each user for each month. These load profiles are representative of the user's average behavior during a specific month. Hence, the hourly electric load profile for the entire year can effectively be reconstructed by assuming that the obtained profile is similar to the daily profile of all the days in the month.

### 3.7. Validation

The validation of the results of this study was conducted by using a set of hourly and monthly measured consumption data [32], and by reconstructing the load profiles using the described methods and comparing them with real load profiles. Kohler et al. [39] provided an extensive review of the metrics that are commonly employed to compare predicted or synthetic load profiles with real ones. Several of the metrics proposed in [39] to assess similarity concentrate on the statistical characteristics of real and simulated load profiles, such as the minimum and maximum values, median, standard deviation, and error on the duration curve. Additionally, metrics based on complexity, such as the number of peaks and the fractal dimension, are also suggested.

In this study, the equality between the reconstructed and real load profiles was evaluated using the Normalized RMSE (NRMSE) and the Normalized MAE (NMAE) for both an entire year and the period when there is contemporaneity between the aggregate consumption and the production of the aggregate power plants ($NMAE_{SE}$, $NRMSE_{SE}$), through Eqs. (12)–(15).

$$NMAE = \frac{\sum_{i=1}^{n} \frac{|Q_i - \widehat{Q}_i|}{n}}{\overline{Q}_i} \tag{12}$$

$$NMAE_{SE} = \frac{\sum_{i=1}^{N} \frac{|Q_i^* - \widehat{Q}_i^*|}{N}}{\overline{Q}_i^*} \tag{13}$$

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n}(Q_i - \widehat{Q}_i)^2}{n}}}{\overline{Q}_i} \tag{14}$$

$$NRMSE_{SE} = \frac{\sqrt{\frac{\sum_{i=1}^{n}\left(Q_i^* - \widehat{Q}_i^*\right)^2}{n}}}{\overline{Q}_i^*} \tag{15}$$

where:

- $Q_i$ represents the real aggregate consumption.
- $\widehat{Q}_i$ represents the simulated aggregate consumption.
- $Q_i^*$ represents the real aggregate consumption during contemporaneity between the aggregate consumption and energy production.
- $\widehat{Q}_i^*$ represents the simulated aggregate consumption during contemporaneity between the aggregate consumption and energy production.
- $n$ is the number of observations.

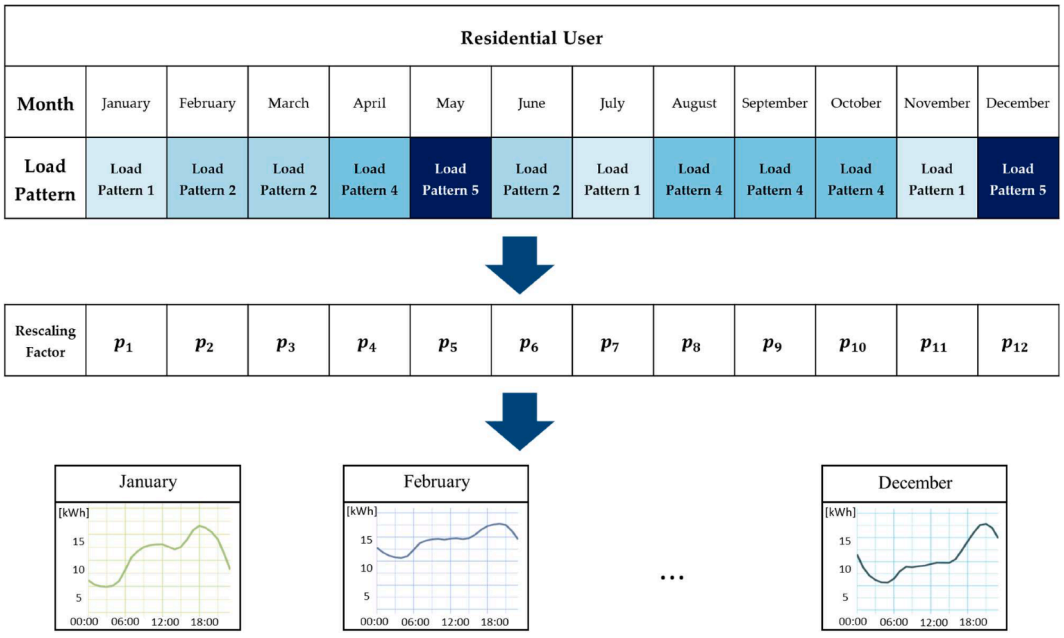



**Fig. 15.** Application scheme of the scaling coefficient process.

The evaluation of NMAE and NRMSE during the period when there is contemporaneity between the aggregate consumption and the production of the aggregate power plants allows to assess of the effectiveness of the proposed framework in reconstructing the aggregate electrical load during the hours of interest for the calculation of shared energy. This aspect is closely related to the estimation of shared energy, which is the main focus of the present research. Moreover, the mean relative absolute error obtained when estimating the shared energy on both a monthly (MRAE) and annual basis (RAE) is evaluated through Eqs. (16) and (17).

$$MRAE = \frac{\sum_{m=1}^{M} \frac{|SE_m - \widehat{SE_m}|}{SE_m}}{M} \tag{16}$$

$$RAE = \frac{|SE_y - \widehat{SE_y}|}{SE_y} \tag{17}$$

where:

- $SE_m$ represents the real shared energy for the m-*th* month.
- $\widehat{SE_m}$ represents the simulated shared energy for the m-*th* month.
- $SE_y$ is the real shared energy for the y-*th* year.
- $\widehat{SE_y}$ is the simulated shared energy for the y-*th* year.
- $M$ is the total number of months.

The shared energy is evaluated, on an hourly, monthly, and annual basis, through Eqs. (18)–(20).

$$SE_h = min\left(E_{cons,h}; E_{prod,h}\right) \tag{18}$$

$$SE_m = \sum_{i=1}^{H} SE_i \tag{19}$$

$$SE_y = \sum_{i=1}^{M} SE_m \tag{20}$$

where:

- $E_{cons,h}$ is the aggregate consumption for the h-*th* hour.
- $E_{prod,h}$ is the energy production of the power plants related to the aggregate consumption for the h-*th* hour.
- $SE_h$ represents the shared energy for the h-*th* hour.
- $SE_y$ represents the shared energy for the y-*th* year.
- $H$ is the total number of hours in a month.

## 4. Case study application results

This section presents the results of applying the proposed methodology to the REC case study, which consists of 37 users randomly chosen from the dataset described in Section 2. The users were selected from a subset of the dataset that excluded the users used during the training phase of the classification step to avoid an unrealistic performance evaluation. The power plant associated with REC consists of a 70 $kW_p$ ground-mounted photovoltaic system in the outskirts of London. The producibility is simulated, on an hourly basis, using PVGIS [40] and the tilt angle and orientation of the photovoltaic panels are calculated using the same tool in order to maximize the system's production, which are set at 40 and −5° respectively. Specifically, the orientation (or azimuth) is the angle of the PV modules relative to the direction due South (−90° is East, 0° is South, and 90° is West). PVGIS is a simulation tool that incorporates solar irradiation and temperature data to accurately simulate the producibility of photovoltaic systems.

Before evaluating the performance of the proposed methodology, in terms of its ability to reconstruct the aggregate electrical load profile, a

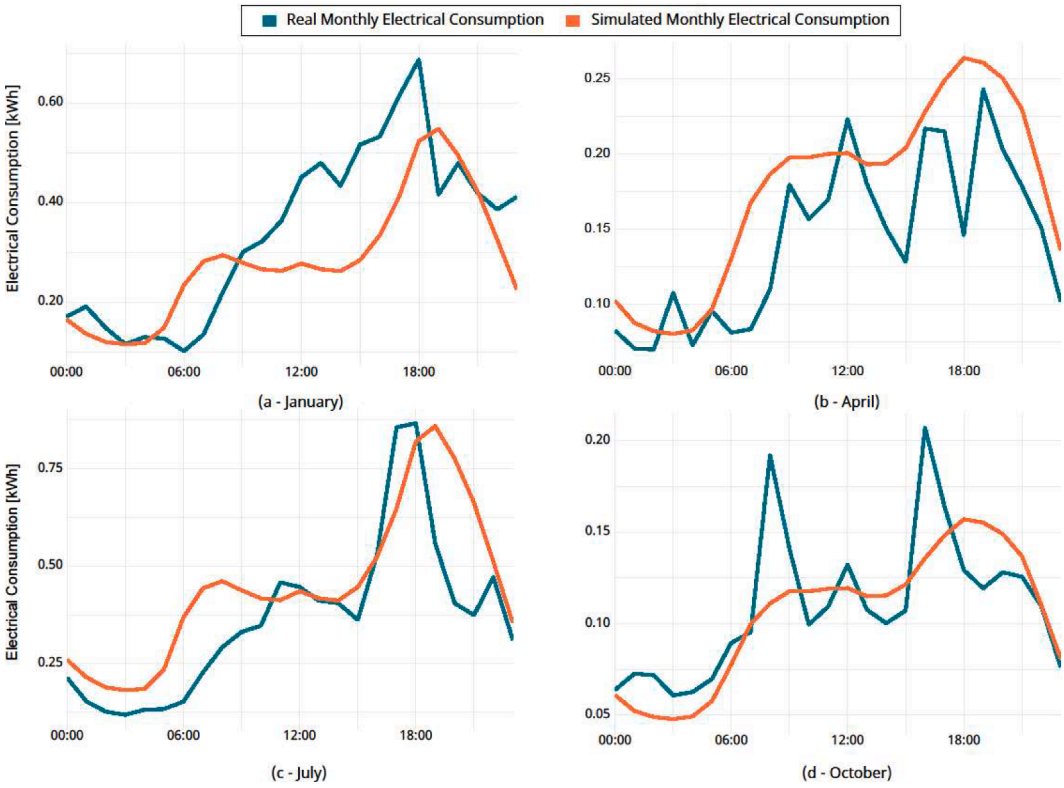



**Fig. 16.** (a) Comparison between the simulated and real monthly average electricity consumptions at the single user level, for the month of January, for one user of the REC; (b) comparison between the simulated and real monthly average electricity consumptions at the single user level, for the month of April, for one user of the REC; (c) comparison between the simulated and real monthly average electricity consumptions at the single user level, for the month of July, for one user of the REC; (d) comparison between the simulated and real monthly average electricity consumptions at the single user level, for the month of October, for one user of the REC.

qualitative comparison between the real and simulated electricity profiles was made at the individual user level for a subset of the users considered in the case study. In particular, the reconstructed load profiles pertaining to four different users are shown in Fig. 16a–d, and compared with the reconstructed profiles for the months of January, April, July, and October, respectively.

As shown, the real and simulated profiles exhibit large differences in some specific time intervals, mainly related to load peaks. For instance, the load peaks in Fig. 16a and d are significantly underestimated and shifted in the simulated load profiles. Furthermore, certain load peaks in Fig. 16b and c are overestimated, while in other cases, they are not even present in the simulated load profiles. This is primarily because the normalized centroids obtained through the Load Pattern Recognition process were used as reference profiles during the hourly profile reconstruction process to generate the real electrical load profile of a given magnitude, and they represent the average electric consumptions of a specific statistical sample. Therefore, the reconstructed load profiles may not be representative of individual users, but rather of the average behavior of a certain number of users. Indeed, the aim of the proposed methodology has been to accurately reconstruct the hourly profile at the aggregate level. Therefore, the use of the normalized centroids as a reference for the aggregate profile takes on a more coherent meaning for the objective of this research. In fact, the ability of the model to reconstruct the load profile of the aggregate, i.e., the sum of all the 37 users, can be appreciated in Fig. 17a–d, where the simulated load curve is compared with the real one, for the months of January, April, July, and December, respectively, and shows that the simulated aggregate load profile is similar to the real one.

By using the typological curves obtained from the load profiles recognition process, which are the centroids of the clusters obtained during that process, we obtain load profiles that represent the average behavior of the users, rather than of individual users, as previously specified. The results obtained from the comparison between the real aggregate profile and the simulated one, using the metrics defined in Section 3.7, are summarized in Table 11.

**Table 11**
Validation metrics evaluated by comparing the simulated load profile with the real one.

| Parameter | Value |
|---|---|
| NMAE | 20.04 % |
| $NMAE_{SE}$ | 18.34 % |
| NRSME | 26.17 % |
| $NRSME_{SE}$ | 23.87 % |
| MRAE | 8.31 % |
| RAE | 0.12 % |

An NMAE of 20.04 %, and an $NMAE_{SE}$ of 18.34 % are obtained, as well as an NRMSE and an $NRMSE_{SE}$ of 26.17 % and 23.87 %, respectively, thus demonstrating that the proposed methodology can be used to more accurately reconstruct the aggregated load profile in the hours where there is simultaneity between the REC consumption and the production of the power plants. This can primarily be attributed to the structure of the rescaling process, where the main factor that influences the transition from typical load profiles to real load curves is the monthly consumption over the "Red" energy consumption range, which describes the energy consumption during the central hours of the day.

The error in calculating the shared energy on a monthly scale (MRAE) and on a yearly one (RAE) is equal to 8.31 % and 0.12 %, respectively, thus demonstrating that the reconstructed profile is able to accurately estimate the shared energy on a monthly basis. The variation of the RAE evaluated for each month, used to evaluate the MRAE, is shown in Table 12.

It can be observed that the model is better at estimating the shared energy, on a monthly basis, during the summer months, while it exhibits a higher relative error in the remaining seasons. This can primarily be attributed to the nature of the residential users in the considered case study. Indeed, in this specific case, it is evident that the electrical load profiles of the case study users are more dependent on consumption over the "Red" energy consumption range during summer months. This

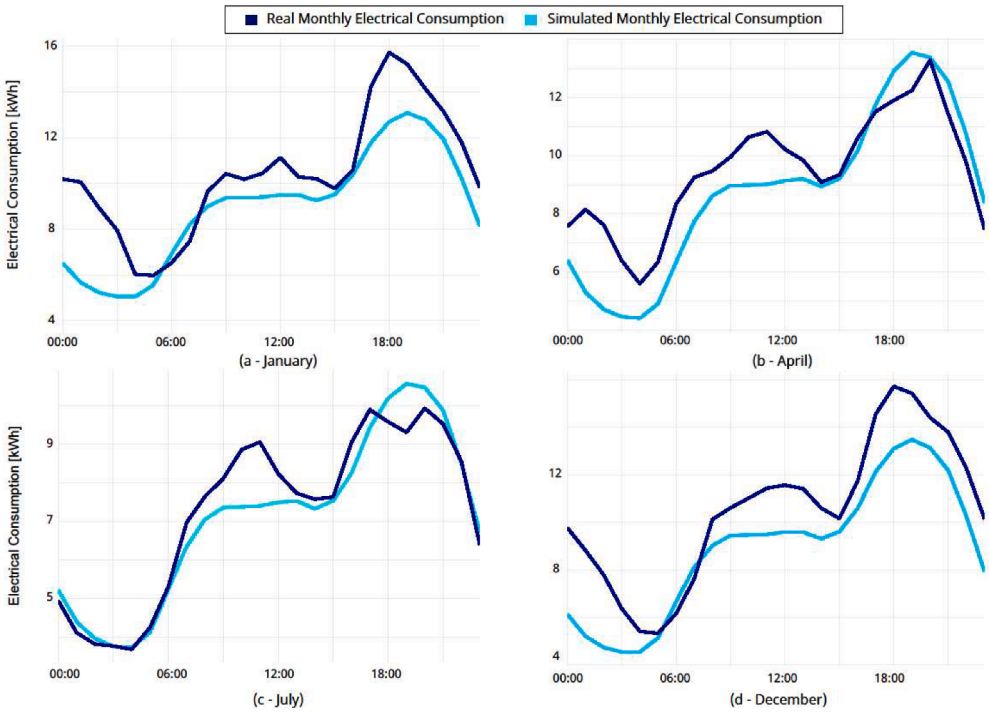



**Fig. 17.** (a) Comparison between the simulated and real monthly average electricity consumptions at the aggregate level, for the month of January; (b) comparison between the simulated and real monthly average electricity consumptions at the aggregate level, for the month of April; (c) comparison between the simulated and real monthly average electricity consumptions at the aggregate level, for the month of July; (d) comparison between the simulated and real monthly average electricity consumptions at the aggregate level, for the month of December.

**Table 12**
The normalized curve values used to calculate the weights for monthly consumption scaling.

| Parameter | Month | Value |
|---|---|---|
| $RAE_m$ | January | 13.32 % |
| | February | 12.83 % |
| | April | 5.84 % |
| | May | 1.05 % |
| | June | 2.27 % |
| | July | 5.62 % |
| | August | 3.14 % |
| | September | 8.48 % |
| | October | 18.06 % |
| | November | 11.38 % |
| | December | 9.46 % |

dependence is less pronounced in the other seasons, with a few exceptions, such as in April and May, where similar RAEs are obtained to the summer period. The similarity can be attributed to the proximity to the summer season, which indicates a transition toward a consumption pattern that is more closely connected to the "Red" energy consumption range. Finally, the relative error obtained when estimating the shared energy on an annual basis is equal to 0.12 %, and it is significantly lower than that of the MRAE. This may be because forecasting errors and deviations between the simulated and real profiles may be offset over a prolonged period, such as a year. This means that negative discrepancies in certain periods could be balanced by positive discrepancies in other periods, thereby resulting in an overall lower relative error for the entire year.

## 5. Discussion

As mentioned in the introduction of this work, most of the previous works on load profiling rely on high temporal resolution data, enriched with additional information on appliances, number of occupants, their characteristics and habits. In contrast, this research proposes a methodology that does not rely on such information-rich datasets, which are often difficult to obtain in real-world contexts. Furthermore, this research aims to apply load profiling techniques in the context of RECs, with the aim of obtaining more accurate estimates of their primary energy parameter, namely the shared energy. This topic represents an emerging novelty in research, in contrast to the numerous load profiling methodologies that focus on reconstructing profiles with high temporal resolution for individual users. Additionally, the framework could be even used in countries where smart meter deployment has already reached a high penetration rate [21] to address the lack of publicly available and accessible data that could be used to improve the estimation of energy flows and, therefore, the REC economic assessments.

The obtained results show that the proposed methodology is capable of reconstructing the electrical load profile of an aggregate from the monthly electrical consumption data of individual users. In general, the model could be trained using datasets from different geographical areas, and could be therefore able to learn information about typical load profiles and their relationship to electrical consumption in different ToU rates in a flexible manner. In the specific case study of this research, the LCL project dataset was used primarily due to the large amount of data that it contains [32]. Using a different dataset, such as a dataset from another country, would likely yield different results. This would start at the load pattern recognition stage, where it is likely that different electrical load shapes would be identified due to different user habits, distribution and use of appliances, or building characteristics. In addition, the relationships between consumption ratios in different ToU rates and load shapes would likely be different, leading to different results in the classification phase. This could result in a different set of important variables for the classification process than those obtained for the specific case study in this article. It is also worth noting that the performance of the classifier is highly dependent on the number of typical profiles to classify, their shapes and their relationship to the consumption values in the different ToU rates. It would be a challenge for the model to classify different load shapes based on very similar electrical consumption values. On the contrary, the classification process would be straightforward if different load shapes corresponded to significantly different electrical consumption values. In the first scenario, it might be necessary to implement a different classification model than the one used in this study. In the second scenario, the model used could potentially improve the classification performance due to the strong relationship that could emerge between the new shape of load profiles and ToU rates.

The increase in the number of typological load profiles could also occur in the light of the electrification of consumption predicted for the coming years [41–44]. Indeed, the use of electric heating and electric vehicles could significantly change the electrical load curves of residential users. For example, the introduction of electric vehicles could generate additional electricity demand during the night hours, effectively changing the shape of the load profile during this time window. However, the impact of these new technologies on the load profile of residential consumers is highly dependent on the characteristics of the installations (i.e., the presence of storage systems linked to photovoltaic plants) and the subsequent use of these technologies by consumers in relation to their habits, making it difficult to predict. In general, it is plausible to expect that the process of electrification of consumption would increase the number of typical load curves that would emerge during the load pattern recognition phase, subsequently causing the classifier to improve or worsen its performance according to the two previous scenarios.

In real-world scenarios, in which hourly or sub-hourly data are usually not available, one of the most commonly used methods for estimating the shared energy generated by users in a REC is to use standard electrical load profiles, often based from statistical analyses at the national scale. These profiles, which represent the average behavior of a large number of users, often do not accurately represent users in more localized geographical areas. As RECs are local aggregations of users and therefore not geographically extensive, it is plausible to expect that the electrical load of users within a REC will generally be different (in some cases very different) from the standard load profiles. Consequently, the use of such profiles, which in most cases consist of one or two profiles per user category (e.g., residential, industrial), would lead to significant errors in the estimation of shared energy. In contrast, the proposed methodology uses typological profiles (as seen in the load pattern recognition phase) that are representative of a very limited number of users, specifically those directly belonging to the REC. Therefore, the typological profiles identified by the model are highly representative of the user behavior within the REC, as well as being more numerous than the standard profiles obtained through statistical analysis. The proposed methodology therefore further refines the estimation of shared energy by using typological load curves that are highly representative of users within the REC and more numerous than standard profiles in different countries.

In the existing literature, many models perform load profiling based on datasets with high temporal resolution and additional information, as previously stated. In addition, such models often focus on reconstructing the electrical load of an individual user rather than an aggregate. Among the reviewed works, the study conducted by Lazzeroni et al. reconstructs the hourly profile of individual users based on monthly electricity consumption data [31]. The purpose of this research is therefore similar to that of the present article, but with a different objective, namely to reconstruct the electrical load for a single user through similar data to those used in our research. The results they obtained, at the individual user level, resulted in a NMAE of approximately 26 %. The proposed methodology, at the aggregate level, achieves a lower NMAE of 20 % when analyzing the profiles for an entire year and 18 % during contemporaneity between the aggregate consumption and production of

the power plants, demonstrating competitive performance at the aggregate level in terms of load profiling,

## 6. Conclusions and further works

In this paper, we present a comprehensive approach to address the lack of high-resolution consumption data for REC members during the "design" phase of an REC project, where a feasibility study and simulation are required to assess the feasibility of the project and to define the optimal design of the REC, that is, to propose a methodology that can be employed by potential aggregators to accurately assess the energy performance of an REC during the initial evaluation phase. We have developed a non-intrusive approach to reconstruct the hourly electrical load profiles of residential users in order to estimate the energy shared within RECs. Our approach involved extensive data preprocessing and the development of a load pattern recognition process to identify typical electrical load profiles. We then created a classification tool that is able to assign the shape of load profiles using predictive attributes extracted from monthly energy bills. We reconstructed the hourly electrical load profiles for an REC case study using a data-driven rescaling process that is based on evaluating the scaling coefficients associated with the most energy-intensive timeframe. Our analysis revealed an NMAE and an NRMSE between the simulated aggregate load profile and the real profile of approximately 20 % and 26 %, respectively. Moreover, we calculated the NMAE and the NRMSE during contemporaneity between the aggregate consumption and production of the power plants, and obtained values of about 18 % and 23 %. In addition, we assessed the ability of the method to estimate shared energy within an REC and obtained MRAE and RAE values of 8.31 % and 0.12 %, respectively. These results demonstrate the viability of our methodological framework, which can be used to estimate shared energy within RECs, even in those data-scarce scenarios that are encountered during the "design" phase of an REC project. Moving forward, our potential future research endeavors will include leveraging on diverse datasets, both public and private, demonstrating the applicability of the proposed framework in different contexts, exploring alternative classification models and rescaling methodologies, and evaluating the errors that occur when estimating shared energy relative to the size and characteristics of the RECs. Furthermore, extending the proposed methodology to encompass user typologies other than the residential sector holds promise for further advancements in the field of RECs.

## Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors report financial support was provided by Polytechnic of Turin.

## Data availability

Data will be made available on request.

## Acknowledgements

This publication is part of the project "Network 4 Energy Sustainable Transition – NEST", Project code PE0000021, Concession Decree No. 1561 of 11.10.2022 adopted by Ministero dell'Università e della Ricerca (MUR), CUP E13C22001890001. Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of 15.03.2022 of MUR; funded by the European Union – NextGenerationEU.

This study was carried out within the Ministerial Decree no. 1062/2021 and received funding from the FSE REACT-EU - PON Ricerca e Innovazione 2014–2020. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

## References

[1] Dóci G, Vasileiadou E, Petersen AC. Exploring the transition potential of renewable energy communities. Futures 2015;66:85–95. https://doi.org/10.1016/j.futures.2015.01.002.

[2] Lowitzsch J, Hoicka CE, Van Tulder FJ. Renewable energy communities under the 2019 European clean energy package – governance model for the energy clusters of the future? Renew Sustain Energy Rev 2020;122:109489. https://doi.org/10.1016/j.rser.2019.109489.

[3] Hanke F, Guyet R, Feenstra M. Do renewable energy communities deliver energy justice? Exploring insights from 71 European cases. Energy Res Soc Sci 2021;80:102244. https://doi.org/10.1016/j.erss.2021.102244.

[4] Minuto FD, Lanzini A, Giannuzzo L, Borchiellini R. Digital platforms for renewable energy communities Projects: an overview. Int J Sustain Dev Plann 2022;17:2007–13. https://doi.org/10.18280/ijsdp.17070.

[5] Heldeweg MA, Saintier S. Renewable energy communities as 'socio-legal institutions': a normative frame for energy decentralization? Renew Sustain Energy Rev 2020;119:109518. https://doi.org/10.1016/j.rser.2019.109518.

[6] Caramizaru, A., Uihlein, A. Energy communities: an overview of energy and social innovation. *Luxembourg: publications office of the European Union* 2020, 30083. https://data.europa.eu/doi/10.2760/180576.

[7] Gjorgievski V, Cundeva S, Georghiou GE. Social arrangements, technical designs and impacts of energy communities: a review. Renew Energy 2021;169:1138–56. https://doi.org/10.1016/j.renene.2021.01.078.

[8] Hoicka CE, Lowitzsch J, Brisbois MC, Kumar A, Camargo LEA. Implementing a just renewable energy transition: policy advice for transposing the new European rules for renewable energy communities. Energy Policy 2021;156:112435. https://doi.org/10.1016/j.enpol.2021.112435.

[9] Bashi MH, De Tommasi L, Cam AL, Relaño LS, Lyons P, Mundó J, et al. A review and mapping exercise of energy community regulatory challenges in European member states based on a survey of collective energy actors. Renew Sustain Energy Rev 2023;172:113055. https://doi.org/10.1016/j.rser.2022.113055.

[10] Volpato G, Carraro G, Cont M, Danieli P, Rech S, Lazzaretto A. General guidelines for the optimal economic aggregation of prosumers in energy communities. Energy 2022;258:124800. https://doi.org/10.1016/j.energy.2022.124800.

[11] Weckesser T, Dominković DF, Blomgren EMV, Schledorn A, Madsen H. Renewable energy communities: optimal sizing and distribution grid impact of photo-voltaics and battery storage. Appl Energy 2021;301:117408. https://doi.org/10.1016/j.apenergy.2021.117408.

[12] Bartolini A, Carducci F, Muñoz C, Kyrki V. Energy storage and multi energy systems in local energy communities with high renewable energy penetration. Renew Energy 2020;159:595–609. https://doi.org/10.1016/j.renene.2020.05.131.

[13] Bianchi FR, Bosio B, Conte F, Massucco S, Mosaico G, Natrella G, et al. Modelling and optimal management of renewable energy communities using reversible solid oxide cells. Appl Energy 2023;334:120657. https://doi.org/10.1016/j.apenergy.2023.120657.

[14] Fioriti D, Frangioni A, Poli D. Optimal sizing of energy communities with fair revenue sharing and exit clauses: value, role and business model of aggregators and users. Appl Energy 2021;299:117328. https://doi.org/10.1016/j.apenergy.2021.117328.

[15] Minuto F, Lanzini A. Energy-sharing mechanisms for energy community members under different asset ownership schemes and user demand profiles. Renew Sustain Energy Rev 2022;168:112859. https://doi.org/10.1016/j.rser.2022.112859.

[16] Garavaso P, Bignucolo F, Vivian J, Alessio G, De Carli M. Optimal planning and operation of a residential energy community under shared electricity incentives. Energies 2021;14:2045. https://doi.org/10.3390/en14082045.

[17] Mignoni N, Scarabaggio P, Carli R, Dotoli M. Control frameworks for transactive energy storage services in energy communities. Control Eng Pract 2023;130:105364. https://doi.org/10.1016/j.conengprac.2022.105364.

[18] Conte F, D'Antoni F, Natrella G, Merone M. A new hybrid AI optimal management method for renewable energy communities. Energy AI 2022;10:100197. https://doi.org/10.1016/j.egyai.2022.100197.

[19] Zhou Y. Advances of machine learning in multi-energy district communities–mechanisms, applications and perspectives. Energy AI 2022;10:100187. https://doi.org/10.1016/j.egyai.2022.100187.

[20] Liu Z, Ma J, Xing C, Liu J, He Y, Zhou Y, et al. Artificial intelligence powered large-scale renewable integrations in multi-energy systems for carbon neutrality transition: challenges and future perspectives. Energy AI 2022;10:100195. https://doi.org/10.1016/j.egyai.2022.100195.

[21] Benchmarking smart metering deployment in the EU-27 with a focus on electricity https://ses.jrc.ec.europa.eu/publications-list/benchmarking-smart-metering-deployment-eu-27-focus-electricity (accessed Oct 23, 2023).

[22] CEER market monitoring report (MMR) https://www.acer.europa.eu/electricity/market-monitoring-report (accessed Oct 23, 2023).

[23] Park JY, Yang X, Miller C, Arjunan P, Nagy Z. Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. Appl Energy 2019;236:1280–95. https://doi.org/10.1016/j.apenergy.2018.12.025.

[24] Pérez-Chacón R, Luna-Romera JM, Troncoso A, Martínez-Álvarez F, Riquelme JC. Big data analytics for discovering electricity consumption patterns in smart cities. Energies 2018;11:683. https://doi.org/10.3390/en11030683.

[25] Wang Q-C, Ding Y, Kong X, Tian Z, Xu L, He Q. Load pattern recognition based optimization method for energy flexibility in office buildings. Energy 2022;254:124475. https://doi.org/10.1016/j.energy.2022.124475.
[26] Fang M, Xiang Y, Xu B-H, Wang T, Pan L, Liu J, et al. Data-driven load pattern identification based on R-vine copula and random forest method. IEEE Trans Ind Appl 2022;58:7919–29. https://doi.org/10.1109/tia.2022.3200920.
[27] Pérez-Ortiz M, Jiménez-Fernández S, Gutiérrez PA, Alexandre E, Hervás-Martínez C, Salcedo-Sanz S. A review of classification problems and algorithms in renewable energy applications. Energies 2016;9:607. https://doi.org/10.3390/en9080607.
[28] Neelamegam S, Ramaraj E. Classification algorithm in data mining: an overview. *Int J P2P Netw Trends Technol* (IJPTT) 2013;4.8:369–74. https://scholar.google.com/scholar?q=Classification%20algorithm%20in%20Data%20mining%20:%20An%20Overview.
[29] Alrawi O, Bayram IS, Al-Ghamdi SG, Koç M. High-resolution household load profiling and evaluation of rooftop PV systems in selected houses in Qatar. Energies 2019;12:3876. https://doi.org/10.3390/en12203876.
[30] Piscitelli MS, Brandi S, Capozzoli A. Recognition and classification of typical load profiles in buildings with non-intrusive learning approach. Appl Energy 2019;255:113727. https://doi.org/10.1016/j.apenergy.2019.113727.
[31] Lazzeroni P, Lorenti G, Repetto M. A data-driven approach to predict hourly load profiles from time-of-use electricity bills. IEEE Access 2023;11:60501–15. https://doi.org/10.1109/ACCESS.2023.3286020.
[32] Schofield J.T., Carmichael R., Tindemans S.H., Bilton M., Woolf M., Strbac G. Low carbon London project: data from the dynamic time-of-use electricity pricing trial, 2013 2016. https://doi.org/10.5255/ukda-sn-7857-2.
[33] Dobre C, Xhafa F. Pervasive computing: next generation platforms for intelligent data collection. Morgan Kaufmann; 2016.
[34] Walfish S. A review of statistical outlier methods. Pharm Technol 2006;30(11):82. https://scholar.google.com/scholar?q=A%20review%20of%20statistical%20outlier%20methods.
[35] National statistics - quarterly energy prices: December 2014 [ARCHIVED CONTENT] (nationalarchives.gov.uk).
[36] Davies DB, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1979:224–7. https://doi.org/10.1109/tpami.1979.4766909. PAMI-1.
[37] Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. Appl Stat 1979;28:100. https://doi.org/10.2307/2346830.
[38] Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory 1982;28:129–37. https://doi.org/10.1109/tit.1982.1056489.
[39] Köhler S, Rongstock R, Hein M, Eicker U. Similarity measures and comparison methods for residential electricity load profiles. Energy Build 2022;271:112327. https://doi.org/10.1016/j.enbuild.2022.112327.
[40] EU Science Hub. PVGIS user's manual. https://ec.europa.eu/jrc/en/PVGIS/docs/usermanual. [Accessed 3 April 2020].
[41] Grottera C, Barbier C, Sanches-Pereira A, Abreu MWde, Uchôa C, Tudeschini LG, Cayla J-M, Nadaud F, Pereira Jr AO, Cohen C, Coelho ST. Linking electricity consumption of home appliances and standard of living: a comparison between Brazilian and French households. Renew Sustain Energy Rev 2018;94:877–88. https://doi.org/10.1016/j.rser.2018.06.063.
[42] Fischer D, Härtl A, Wille-Haussmann B. Model for electric load profiles with high time resolution for German households. Energy Build 2015;92:170–9. https://doi.org/10.1016/j.enbuild.2015.01.058.
[43] Besagni G, Premoli Vilà L, Borgarello M. Italian household load profiles: a monitoring campaign. Buildings 2020;10:217. https://doi.org/10.3390/buildings10120217.
[44] Zapata Castillo V, De Boer H, Maícas Muñoz R, Gernaat DEHJ, Benders R, van Vuuren D. Future global electricity demand load curves. SSRN Electron J 2021. https://doi.org/10.1016/j.energy.2022.124741.