A scalable ensemble approach to forecast the electricity consumption of households

Lola Botman, Jonas Soenen, Konstantinos Theodorakos, Aras Yurtman, Jessa Bekker, Koen Vanthournout, Hendrik Blockeel, Bart De Moor *Fellow, IEEE & SIAM* and Jesus Lago

***Abstract*—Long-term individual household forecasting is useful in various applications, e.g., to determine customers' advance payments. However, the literature on this type of forecasting is limited; existing methods either focus on short-term predictions for individual households, or long-term prediction at an aggregated level (e.g. neighborhood). To fill this gap, we present a method that predicts the monthly consumption of individual households over the next year, given only a few months of consumption data during the current year. Utility companies can exploit this method to predict the consumption of any customer for the next year even with incomplete data. The method consists of three steps: clustering the data using *k*-means, prediction using an ensemble of forecasts based on the historical median distribution among similar households, and smoothing the predictions to remove weather-dependent patterns. The method is highly accurate as it finished third in the IEEE-CIS competition (and ranks first when leveraging insights from another team), focused on forecasting long-term household consumption with incomplete data. It is also very scalable thanks to its low computational complexity and weak data requirements: the method only requires a few months of historical data and no household-specific or weather information.**

***Index Terms*—Long-Term Load Forecasting, Clustering, Smart Meter, Household Consumption.**

I. Introduction

ELECTRICITY forecasting has been studied for many years and remains an important task. The roll-out of domestic smart meters enables the collection of consumption data of individual households. Electricity forecasting is one way in which these data can be used to create value. Accurate electricity consumption prediction is useful for a plethora of applications including, but not limited to, determining customers' advance payments, supporting day-to-day grid operations and strategic planning of energy grid extensions [1]. Different applications require different types of forecasts: day-to-day grid operations might require an hourly forecast for the next day while strategic planning might require a monthly forecast for the next ten years [2]. In this paper, we propose a novel algorithm to forecast the monthly electricity consumption of individual households for the next year, given the monthly consumption during the available months of the current year.

The algorithm is one of the winning approaches at the IEEE-CIS competition [3]. The goal was to forecast the monthly consumption of multiple British households in 2018 using smart meter data (electrical consumption), weather data from 2017, and additional household-specific information such as the number of occupants, the number of bedrooms, etc. In the original competition, our approach ranked third out of 71 participating teams, reaching state-of-the-art accuracy, and we show that our method outperforms the approaches ranked first and second when disregarding smart meters without battery or when applying the same post-processing trick that the first approache applied to those faulty smart meters.

A key asset of the algorithm is that it can make predictions for a full year ahead even with incomplete data, e.g., even for households with one month of historical data. In addition, the proposed approach has the benefit of requiring low computational power, which grants the ability to scale up across millions of households. Finally, together with the other competition submissions, this model is one of the first to solve the challenge of long-term electricity forecasting at the individual household level in the literature.

A. Related work

The method proposed in this paper forecasts the monthly electricity consumption of individual households one year ahead, requiring only historical data of the previous year, at least the data from one month.

Electricity consumption forecasting is a wide subject of research. It has two dimensions: horizon and spatial granularity. The horizon is the forecast length in the future. It can be from few hours to multiple years. The spatial granularity is a geographical specification, i.e., consumption of an appliance or of an entire a country. Another relevant parameter is the sampling rate, which can vary from one second to one year. It is strongly linked with the horizon, e.g., if the electrical consumption of a city had to be predicted ten years ahead, an hourly forecast would not be accurate and thus not be considered.

Most of the efforts in the literature have been focused on long-term forecasting of coarse granularity, e.g., cities, geographical zones [4], regions [5], [6] or countries [7]–[9]. In long-term forecasting, due to the randomness and high volatility of the individual household electricity consumption,

Preprint re-submitted to IEEE Transactions on Smart Grid on 06/05/2022.

L. Botman, K. Theodorakos, B. De Moor are with the research group STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics within the department of Electrical Engineering (ESAT) (Corresponding author: lola.botman@kuleuven.be).

J. Soenen, A. Yurtman, J. Bekker and H. Blockeel are with the research group of DTAI Declarative Languages and Artificial Intelligence within the department of Computer Science (CS).

K. Vathournout is with VITO/EnergyVille.

Jesus Lago is with Amazon. He contributed to this work as an outside activity and not as part of his role at Amazon.

existing methods rely on averaging and aggregating these consumption patterns to a coarser spatial granularity.

Methods for individual household, i.e., fine granularity, consumption have been proposed [10]–[13]. However, they are limited to short horizons. Typically they forecast a few hours up to several days ahead and only for a small number of households. The latter is either due to the complexity of the algorithm used, leading to high computational requirements [10], or the lack of available input data. For example, in the Convolutional Neural Network - Long Short Term Memory (CNN-LSTM) model in [11], the authors used a configuration (i.e., three smart meters in specific household locations) that is nearly impossible to replicate in thousands of households. Another example of data scarcity is given in [12], where the case study is based on an unoccupied house and all activities are of an academic nature, which reduces stochasticity and simplifies the forecasting problem. The model proposed in [13] suffers the same drawback. It relies on typical daily load profiles depending on three parameters: the number of occupants, the time at which the first person gets up in the morning and the last person goes to sleep and the part of the day during which the house is unoccupied. The influence of the individual appliances is also considered (electric oven, TV, water heating, etc.). As with the previous works, this method can hardly generalize to multiple households due to the type of data required (e.g., disclosing the time of waking up or going to bed might conflict with privacy laws in several countries).

Short-term individual load forecasting might require different evaluation metrics [14] than long-term forecasting, e.g., to take in account the double penalty effect generated by the consumption peaks. Classical error metrics, such as RMSE, penalize twice for a peak that is correctly predicted by the algorithm in terms of amplitude and duration, but displaced in time: once when the peak actually happens and is not predicted and once again when the peak is predicted but it does not actually happen. Long-term forecasting, based on monthly aggregated data for example, are less prone to peaks, thanks to the averaging and are thus not subject to this double penalty effect.

Research on long-term and fine granularity forecasting mostly focuses on large energy consumption units, e.g., supermarkets [15] or public buildings [16], [17], where the patterns are less stochastic than households. Others investigate long-term residential consumption forecasting, per customer type, depending on average income and other demographics [18]. In the latter case, individual household consumption is not predicted and the proposed method requires 30 years of historical consumption data in the training phase.

The lack of research on long-term household forecasting can further be seen in the extensive review of Zhao et al. [19], where out of the 92 extensively reviewed papers, only eight tackle residential load forecasting and just two consider fine granularity residential monthly forecasting. The first proposed method [20] tackles individual residential forecasting by disaggregating the household load into sixteen appliance categories, using detailed information such as the historical consumption, energy prices, weather information and demographics obtained through a questionnaire. Although potentially accurate, this method is not scalable due to required questionnaire data. The second one [21] proposes an approach in three steps in order to predict monthly consumption of six family houses based on only one month of historical data of the heating demand and the domestic energy demand as well as the indoor-outdoor temperature difference. The first step is to generate the consumption of a reference building using a simulation tool based on blueprint data of a similar house, the second is to generate additional data by scaling the measured data, the third is to train a neural network on the extended data and finally make predictions based on the actual indoor-outdoor temperature and the reference building. Several limitations are to be noted. Firstly, detailed blueprint data of a similar house is necessary, secondly the total household consumption has to be disaggregated into the energy used for space heating and the energy used for the electric appliances and hot water and finally, the data augmentation using scaling only makes sense because of the location of the households under investigation. Indeed, the houses are in Umeå, Sweden, where the temperature varies between $-30\,^{\circ}\mathrm{C}$ and $30\,^{\circ}\mathrm{C}$.

B. Motivation and contributions

This paper aims to fill a gap in the literature regarding long-term individual household forecasting, proposing a highly accurate and scalable method based on relative consumption patterns and ensemble learning. The proposed method is the first of its kind, as it

- can forecast *long-term* consumption of *individual* households,
- can operate on households with incomplete data,
- can handle missing time samples,
- does not have complex and exogenous data requirements such as weather data or household-specific attributes.

The method is highly accurate (as demonstrated by the award in the IEEE-CIS competition [3]), has low computational complexity and can be scaled to any number of households. As motivated by the competition, this type of forecasting is beneficial for customers, for accurate bill estimations. It is also interesting for producers, to schedule the right amounts of electricity production and procurement; for suppliers, to detect potential discrepancies between self-declared and real consumption; and for network operator in the grid management and strategic planning. Additionally, long-term individual household forecasting is one of the main research recommendation topics in the context of the energy transition [9].

The proposed method has to overcome a series of challenges that existing methods, due to the nature of the consumption granularity, do not have to overcome. For instance, household consumption patterns are more stochastic than aggregated data (on the city or national level) and also more stochastic than the consumption patterns for larger buildings such as supermarkets or offices. The algorithms that work for a supermarket, offices or a large building might not work for a household. Moreover, households change distributors every year in order to obtain the best rates. A commercial supplier might not have a full year of historical data to predict the next year. It is thus crucial to have a method which is reliable, works for households and can predict longer horizons than the available historical data.

C. Organization of the paper

The paper is organized as follows. Section II presents the key attributes considered to build the model, based on a set of realistic assumptions. The different steps of the method are detailed in Section III, as well as the model parameters. Appendix A analyzes the impact of each step on the final performance metric. The case study proposed by the IEEE Data Port platform is explained extensively in Section IV, including information on the dataset, the benchmark models, the performance metrics and the results. Finally the outcome and perspectives are discussed in Section V.

II. Key attributes of the model

Before explaining the details of the model (see Section III), it is important to outline the key attributes of the model and the assumptions made to derive these attributes. The proposed model forecasts monthly electrical consumption of individual households for a full calendar year using incomplete household data from the previous calendar year, where at least one month of measurements must be available. The model does not require household-specific information nor weather data. The methodology presented consists of four key attributes, which are based on assumptions, detailed in this section.

As the granularity of the predictions is monthly, most of the hourly and daily stochastic behavior of individual households averages out. In a monthly resolution, a lot of the stochastic behavior of individual households, e.g., what time one goes to sleep, plays little role. Instead, the factors that drive the consumption at that level are more deterministic, e.g., how large the household is. Since these deterministic factors are likely to be shared across many households, simple methods that use the mean or median consumption of households with similar consumption patterns, should be able to make highly accurate predictions. The first and second key attributes of the model stem from this assumption. The first attribute is that **the proposed model disregards hourly and daily patterns, and directly works with monthly data**. Therefore, even if hourly data from smart meters is available, we aggregate the data to the monthly level and disregards the stochastic patterns of lower resolutions. As a second key attribute, **we make predictions based on the median consumption of *similar* households**.

Although the absolute consumption of households differs significantly, relative monthly consumption patterns have less variance and are more similar. Therefore, the perceptual consumption of each month with respect to the yearly consumption, is shared among many households. This can be explained by the fact that, independently of the household type or size, similar human behavior is shared across the households. Based on this observation, a third key attribute of the model is to **work with relative consumption patterns instead of absolute ones**. In other words, we normalize electricity consumption data.

The model assumes that, since relative consumption patterns are considered, additional household-specific information such as household size and weather data play little to no role. In particular, households with different types of appliances and a different number of occupants might have the same relative consumption profile, while households with the exact same number of occupants and/or appliances might have different relative profiles. Similarly, weather data are only available for the current year. Using this weather data to predict the next year can lead to over-fitting the model to the current year. As an example, let us consider a household where November in the current year is colder than December: that household will likely have a larger consumption in November than December due to the weather. Yet, that specific situation will likely not generalize to the next year as December is, in general, expected to be colder. Moreover, weather forecasts are usually not available one year in advance and can thus not be used for this horizon. For these reasons, the fourth model attribute is to **disregard exogenous features**.

III. Methodology: Ratio approach

In order to forecast the household monthly consumption, different steps are applied sequentially. These can be divided in three main parts. Firstly, the training pipeline, Fig. 1a, consists of the preprocessing, data augmentation, normalization and clustering. Secondly, the inference framework, Fig. 1b, consists of the prediction step. Finally, the ensemble and post processing steps are illustrated in Fig. 1c. All the modules are detailed in Subsections III-A to IV-E. In these sections, we provide a clear example for interpretability, while a general notation is used in the figures.

- *Preprocessing* A real life dataset usually contains missing data, due to defaults in the smart metering device for example. For each household/smart meter, full missing days are imputed using linear interpolation on a daily level. Subsequently, the data are aggregated to the monthly level using summation.
- *Data Augmentation & Normalization* As discussed in Section II, it is better to consider relative as opposed to absolute consumption. This conversion happens in the normalization step. However, a consumption profile depends heavily on the month in which the customer signs up with a supplier and this is reflected in the normalization. Therefore, our proposed method explicitly considers the sign-up month. This is explained in more detail in Section III-A.
- *Clustering* The relative consumption profiles are clustered together in order to group similar profiles. Different clusterings are executed per sign-up month. Profiles that signed up later are retroactively added to clusterings for earlier sign-up months. This is explained in more detail in Section III-B.
- *Prediction* Each cluster can be used to make a prediction for a certain household. This operation is explained in Section III-C. Since multiple clusterings exist per household, several predictions can be made. Section III-D explains which clusters to consider and how to combine their predictions in an ensemble approach.
- *Postprocessing* As a final step, the predictions are smoothed using a standard moving average technique in order to reduce the effect of weather patterns in the

current year, as explained in Section II. By using a moving average, the effects of weather are mitigated as the final prediction is the average between nearby months. This step has a parameter w describing the window length for the moving average.

The entire process makes use of very few hyperparameters. Only three parameters require tuning (the minimum cluster size $n_{\min,\text{cluster}}$, the number of nearest neighbors n_{similar}, and the window length w for smoothing), which are explained in Section III-B, III-C, III-E respectively. As detailed in Section IV-H and Appendix B, their value, although important, is not very critical as the method is rather robust to the selection of these three hyperparameters.

A. Data Augmentation & Normalization: relative consumption profiles

The proposed method is based on using relative consumption profiles, where a relative profile is the absolute profile divided by the yearly consumption. For instance, the relative consumption of a given household in April is the consumption of April divided by the annual consumption.

The vector of monthly consumption of a household x_j in a year is defined as

$$\mathbf{L}^{x_j} = [l_1^{x_j}, \ldots, l_{12}^{x_j}] \in \mathbb{R}^{12} \tag{1}$$

where the subscripts of the vector's elements denote the month, e.g., $l_1^{x_j}$ is the monthly consumption value for January. For a household x_j that signed up in month s, the first $s-1$ elements of the monthly consumption vector $\mathbf{L}^{x_j}$ are missing:

$$\mathbf{L}^{x_j} = [\underbrace{\text{NaN}, \ldots, \text{NaN}}_{(s-1)}, l_s^{x_j}, \ldots, l_{12}^{x_j}] \in \mathbb{R}^{12}. \tag{2}$$

Therefore, we introduce the monthly consumption vector

$$\mathbf{L}_s^{x_j} = \left[l_s^{x_j}, \ldots, l_{12}^{x_j}\right] \in \mathbb{R}^{12-s+1} \tag{3}$$

of household x_j starting from month s. This notation will also be used to discard the first few months of a profile even though consumption data of the discarded months might be available.

To make the consumption profiles between households that signed up in the same month comparable, the consumption profiles $\mathbf{L}_s^{x_j}$ are normalized into relative consumption profiles $\mathbf{R}_s^{x_j}$ as

$$\begin{aligned}\mathbf{R}_s^{x_j} &= \left[r_s^{x_j}, \ldots, r_{12}^{x_j}\right] \in \mathbb{R}^{12-s+1}. \\ &= \frac{1}{\sum_{i=s}^{12} l_i^{x_j}} [l_s^{x_j}, \ldots, l_{12}^{x_j}].\end{aligned} \tag{4}$$

After this transformation the total relative consumption of each household sums up to one.

The relative consumption of households that signed up in different months are still not directly comparable. For example, a household x_j that signed up in November has a relative consumption in November that is around 50%, but if the same household has signed up in January the relative consumption during November would be closer to 8% ($\approx 1/12$).

To solve this issue, two steps are taken. In the clustering step, to find similar consumption patterns, a different clustering is made for each sign-up month. In the prediction step, only ratios between two months are considered as these ratios are independent of the sign-up month (see section III-C for details).

Furthermore, a data augmentation technique is applied to maximally exploit the available data. When making a clustering for relative profiles that signed up in month s, we consider not only the data of households that signed up in month s, but also the data of the households that signed up earlier by discarding any consumption before month s. For example, a profile with sign-up month March, is also considered as if it signed up in April, i.e., by disregarding March. Similarly, it is also appended to the profiles signed up in May, i.e., by disregarding March and April; and so on. The relative profiles of the households that signed up earlier are re-normalized such that each relative profile sums up to one. In the prediction step, a similar procedure is applied, which is explained in section III-C.

B. Clustering: grouping similar profiles

As households from different sign-up months have different relative values, clusters are built for each possible sign-up month. As an example, Fig. 2 illustrates the four clusters obtained when considering the households that signed up in January, using the *k*-means algorithm with euclidean distance metric. Similarly, Fig. 3 illustrates the five clusters obtained when considering the households that signed up in June, including those that signed up before and from which the measurements of the months before June have been ignored. In both figures, for each cluster, the centroid (mean) of the cluster is plotted together with a 90% confidence interval. The *k*-means clustering is guaranteed to converge to an optimum, although this might be a local optimum [22].

In order to introduce the notation for the prediction step (Section III-C), let $x_1, x_2, \ldots, x_N$ be the N households that signed up in month s or earlier. The set of these households is denoted by $\mathcal{X}_s$. For each of these households, independently of the data available, the clustering algorithm first defines a relative profile $\mathbf{R}_s^{x_j} \in \mathbb{R}^{12-s+1}$ starting in s, as in (4). Let $\mathcal{S}_s$ be the set of the N relative profiles starting in s (one per household that signed up in month s or earlier):

$$\mathcal{S}_s = \{\mathbf{R}_s^{x_1}, \ldots, \mathbf{R}_s^{x_N}\}. \tag{5}$$

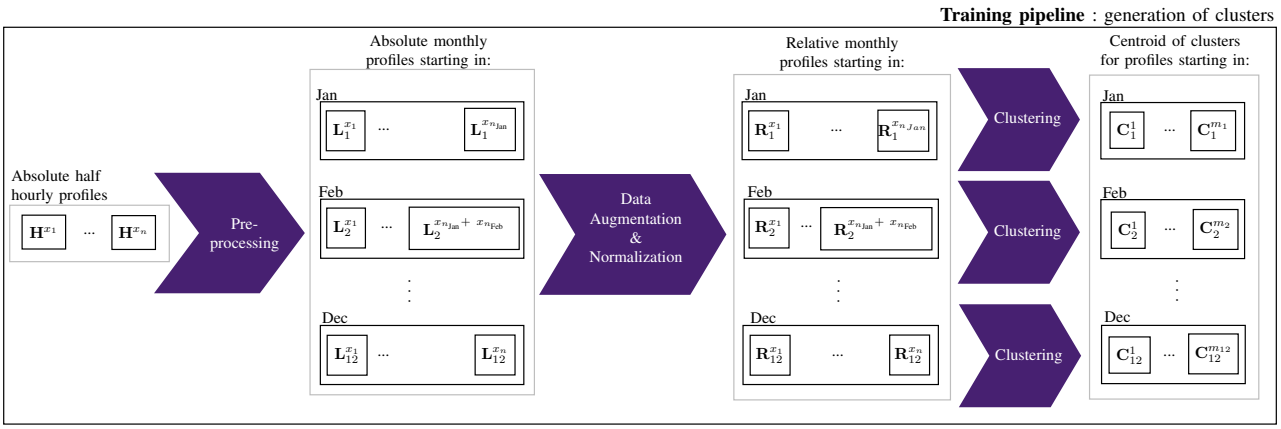
The clustering step uses the K-mean clustering algorithm [23] with euclidean distance to cluster the relative consumption profiles $\mathcal{S}_s$ into a set of m_s clusters with centroids

$$\mathcal{C}_s = \{\mathbf{C}_s^1, \ldots, \mathbf{C}_s^{m_s}\}, \tag{6}$$

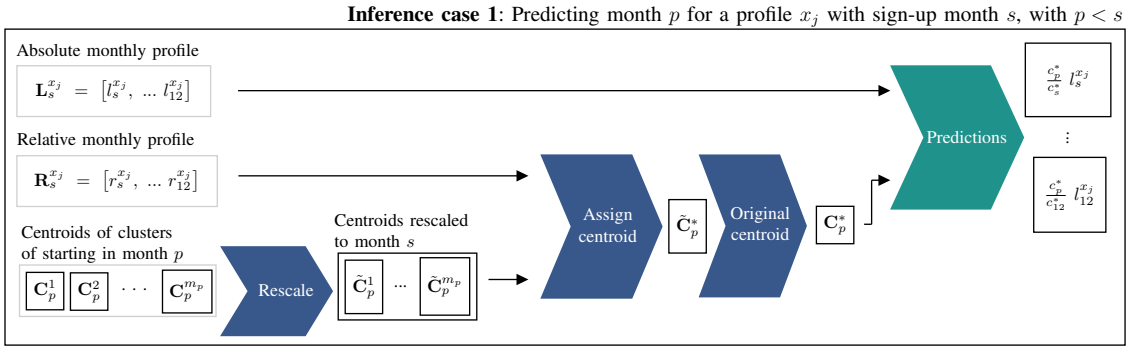
where the centroid of the k'th cluster $\mathbf{C}_s^k$ is the mean of all profiles in that cluster.

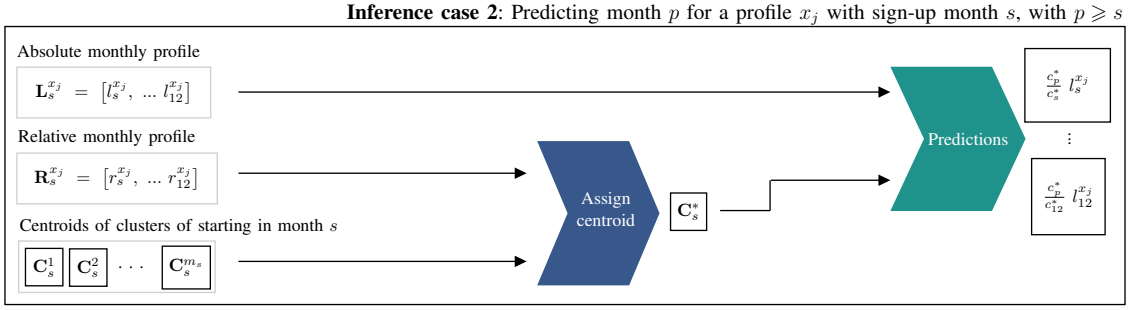
$$\mathbf{C}_s^k = [c_{s,s}^k, \ldots, c_{s,12}^k]. \tag{7}$$

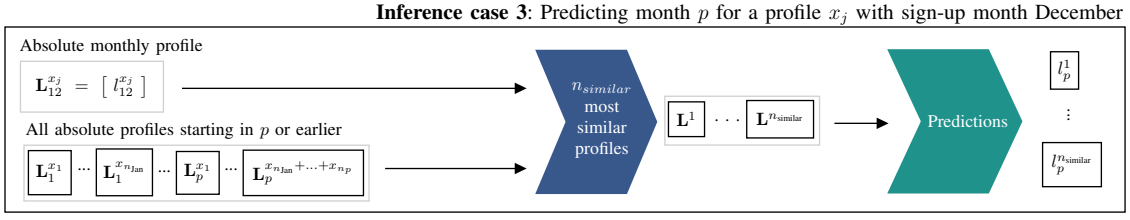
The double subscripts v, w in $c_{v,w}$ represent the sign-up month of the centroid and the given month within the centroid



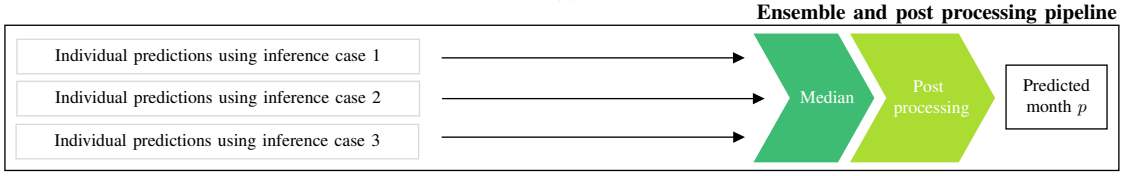
(a)







(b)



(c)

Fig. 1. Flowchart of the method proposed in this paper. The input and output data are in the rectangles and the steps executed are in the chevron arrows. (a) The training pipeline is applied to all the profiles and aims to build the clusters based on the relative profiles. The data augmentation step is highlighted using the indices of the profiles. There are n_{Jan} profiles which start in January, n_{Feb} in February, and so on. The total number of profiles is $n = n_{\text{Jan}} + n_{\text{Feb}} + \cdots + n_{\text{Dec}}$. (b) General framework for inferring the consumption of a month p for a profile with sign-up month s. (c) Ensemble and post processing pipeline: the individual predictions from the inference cases, are combined using the median and smoothed in the post-processing step.

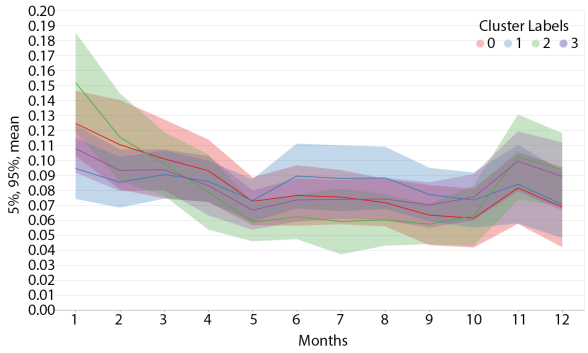


Fig. 2. Example of clusters of relative consumption profiles starting in January for the IEEE-CIS competition data. The number of clusters obtained is four, determined using the elbow method. The centroid (mean) of the cluster is plotted together with a 90% confidence interval.

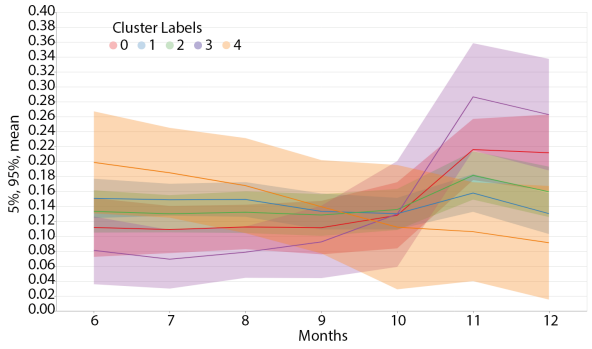


Fig. 3. Example of clusters of relative consumption profiles starting in June for the IEEE-CIS competition data. The number of clusters obtained is five, determined using the elbow method. The centroid (mean) of the cluster is plotted together with a 90% confidence interval. It is clear, when comparing Fig. 2 and Fig. 3 that the relative values are different depending on the sign-up month.

respectively. For example $c_{3,6}$ represents the month June within the centroid C_3, which has sign-up month March.

As *k*-means requires the number of clusters as an input, the elbow method [24] is employed to estimate the number of clusters. Moreover, to avoid unrepresentative clusters that do not generalize, we discard clusters with fewer than $n_{\min,\text{cluster}}$ profiles, as they are likely to contain outliers and may not be representative enough.

C. Prediction: the power of the median via ensemble learning

The proposed model forecasts monthly electrical consumption of individual households for a full calendar year using (incomplete) household data from the previous calendar year. Here, "predicted month" refers to the monthly consumption value that we aim to predict in the next year, while "sign-up month" refers to the month in which the household signed up with the provider, meaning in the previous year.

Once we have clustered the relative profiles, we build the predictions. To do so, for each household to be predicted, we generate multiple predictions based on the clustered profiles and compute the final prediction as the median of these predictions. We combine predictions to exploit ensemble learning [25], i.e., combining predictions (or models) to reduce the bias or variance errors of the individual predictions (or models). In this context, we use the median instead of the mean to combine the predictions because it is less sensitive to outliers.

The proposed method considers a slightly different ensemble learning step depending on whether the prediction is made for an earlier or later month than the month in which the household signed up. In particular, the method distinguishes three cases, as illustrated in Fig. 1b: (1) predicting an earlier month than the sign-up month, i.e., a month for which there is no monthly data of the previous year available, (2) predicting a later (or equal) month to the sign-up month, i.e., a month for which there is monthly data of the previous year available, and (3) predicting the consumption for households that signed up in December.

To explain the prediction step, it is easier to consider an example. Let us consider again a household x_j that signed up in June. As explained in the previous sections, for household x_j, the relative profile is from June onward:

$$\mathbf{R}_6^{x_j} = [\mathbf{r}_6^{x_j}, \ldots, \mathbf{r}_{12}^{x_j}]. \tag{8}$$

1) Predicting an earlier month than the sign-up month: The prediction process for this case is represented in Fig. 1b, case 1. Let us assume that the month of March needs to be predicted for household x_j. For that, consider the centroids $\mathcal{C}_3 = \{\mathbf{C}_3^1, \ldots, \mathbf{C}_3^{m_3}\}$ obtained from the clustering with March as a sign-up month and denote the centroid elements as:

$$\mathbf{C}_3^k = [c_{3,3}^k, \ldots, c_{3,12}^k] \in \mathbb{R}^{10}. \tag{9}$$

Second, based on these m_3 centroids, a set of re-scaled centroids are generated $\{\tilde{\mathbf{C}}_3^1, \ldots, \tilde{\mathbf{C}}_3^{m_3}\}$ which match the scaling and size of the relative profile $\mathbf{R}_6^{x_j}$ of household x_j (with sign-up month June):

$$\begin{aligned}\tilde{\mathbf{C}}_3^k &= \left[\tilde{c}_{3,6}^k, \ldots, \tilde{c}_{3,12}^k\right] \in \mathbb{R}^7 \\ &= \frac{1}{\sum_{i=6}^{12} c_{3,i}^k} \left[c_{3,6}^k, \ldots, c_{3,12}^k\right].\end{aligned} \tag{10}$$

The part of the centroid before the sign-up month is disregarded and the centroid is re-scaled so that the new centroids sum up to one. Third, we determine which centroid $\tilde{C}_3^k$ lies closest to the relative profile $\mathbf{R}_6^{x_j}$, using Euclidean distance. We thus have the optimal centroid

$$\tilde{\mathbf{C}}_3^\star = \underset{\tilde{\mathbf{C}}_3^k \in \{\tilde{\mathbf{C}}_3^1, \ldots, \tilde{\mathbf{C}}_3^{m_3}\}}{\arg\min} \|\tilde{\mathbf{C}}_3^k - \mathbf{R}_6^{x_j}\|_2^2. \tag{11}$$

Then, the original full centroid $\mathbf{C}_3^\star \in \mathbb{R}^{10}$ associated with the optimal rescaled centroid $\tilde{\mathbf{C}}_3^\star \in \mathbb{R}^7$ is used to predict March's consumption for household x_j. Seven predictions are computed, based on the seven relative centroid values $c_{3,3}^\star, \cdots, c_{3,12}^\star$ of $\mathbf{C}_3^\star$ and the seven absolute measurements $l_6^{x_j}, l_7^{,x_j}, \cdots, l_{12}^{x_j}$ of household x_j. The predicted consumption of month March $\hat{l}_3^{x_j}$ is calculated as the median of these individual predictions:

$$\hat{l}_3^{x_j} = \text{median}\left(\frac{c_{3,3}^\star}{c_{3,6}^\star} l_6^{x_j}, \ldots, \frac{c_{3,3}^\star}{c_{3,12}^\star} l_{12}^{x_j}\right). \tag{12}$$

In other words, to predict the consumption of March for household x_j, seven predictions are built: one for each available monthly consumption. In particular, for each available monthly consumption $l_q^{x_j}$, a prediction is built by multiplying the ratio between the representative relative consumption in March and the representative relative consumption in month q, i.e., $\frac{c^\star_{3,3}}{c^\star_{3,q}}$, by the monthly absolute consumption $l_q^{x_j}$. Then, the final prediction is built as the median of the individual predictions.

2) Predicting a later month than the sign-up month: This prediction process is represented in Fig. 1b, case 2. When predicting a later (or equal) month than the sign-up month the same principle applies. However, instead of working with the clusters starting in the month that we want to predict p, the clusters associated with the sign-up month s are used.

As an example, let us consider again a household x_j that signed up in June. However, let us consider the case of predicting July rather than March. Now, instead of using the m_7 centroids $\{\mathbf{C}_7^1, \ldots, \mathbf{C}_7^{m_7}\}$ associated with the predicted month (July), the m_6 centroids $\{\mathbf{C}_6^1, \ldots, \mathbf{C}_6^{m_6}\}$ associated with the sign-up month (June) are used.

The reasons for considering the centroids of the sign-up month instead of the predicted month are twofold. First, as the scale and length of each centroid $\mathbf{C}_6^k \in \mathbb{R}^7$ is the same as the relative profile $\mathbf{R}_6^{x_j} \in \mathbb{R}^7$, it is not necessary to re-scale the centroids. Second, by working with the clusters associated with the sign-up month, the number of predictions used to compute the median is maximized. For example, while for June as a sign-up month there are seven predictions, whereas for July there are only six. Similarly to the previous case, the closest centroid (in terms of Euclidean distance) to the relative profile $\mathbf{R}_6^{x_j}$ is determined and denoted by $\mathbf{C}_6^\star$, as in (11). Similar to (12) in the previous case, the prediction $\hat{l}_7^{x_j}$ is built by computing the medians of the ratios between the representative relative consumption of July and the representative relative monthly consumption of the available months multiplied by the absolute monthly consumption:

$$\hat{l}_7^{x_j} = \text{median}\left(\frac{c^\star_{6,7}}{c^\star_{6,6}} l_6^{x_j}, \frac{c^\star_{6,7}}{c^\star_{6,7}} l_7^{x_j}, \ldots, \frac{c^\star_{6,7}}{c^\star_{6,12}} l_{12}^{x_j}\right). \quad (13)$$

3) Predicting households that signed up in December: As the relative consumption for the households that signed up in December is always 1, the described method cannot be used to predict these households. Therefore, for these households, a different approach is considered, represented on Fig. 1b, case 3.

First, for each household x_j and predicted month p, the n_{similar} most similar households are computed in terms of absolute consumption in December that have data available for month p. That is, to predict March we consider all meters with data for March and December, in other words, the set of households that signed up in month p or earlier.

Then, the n_{similar} meters that are most similar to household x_j are found. Second, defining this set of similar households by $\mathcal{S}^{p,j} = \{x_1^{p,j}, \ldots, x_{n_{\text{similar}}}^{p,j}\}$, the prediction $\hat{l}_p^{x_j}$ is built for household x_j and month p as the median of the historical values of the similar households:

$$\hat{l}_p^{x_j} = \text{median}\left(l_p^{x_1^{p,j}}, \ldots, l_p^{x_{n_{\text{similar}}}^{p,j}}\right). \quad (14)$$

To find the n_{similar} most similar households, the k-nearest neighbors algorithm [26] is applied. In this context, although the parameter n_{similar} should be optimized, empirical observation shows that it makes no difference for values $n_{\text{similar}} > 10$. We have tested values ranging from 10 to 100 during the IEEE-CIS competition and empirically observed that it does not have a significant impact. The value used in the final submission of the method is $n_{\text{similar}} = 50$.

D. Ensemble learning

As explained in the previous sections, the consumption of each (household, month) pair is predicted by building multiple predictions for each pair and then computing the median. The reason for doing so is ensemble learning [25], i.e., combining predictions (or models) to reduce the bias or variance errors of the individual predictions (or models).

In general, when combining multiple predictions, the larger the ensemble, the lower the variance or bias error and the better the prediction becomes. To that end, as an additional step during the prediction phase, a data augmentation technique is performed to increase the number of predictions for each (household, month) pair.

As before, let us explain this step with an example. Let us consider the household x_j that signs up in June and the process of predicting July. As explained in Section III-C2, this prediction is done by computing the median of seven individual predictions (cfr. (13)).

To improve this prediction, additional individual predictions are generated by assuming that the household did not sign up in June but signed up in July. To do so, the exact same procedure described in Section III-C2, is repeated but assuming that the household has no data for June. This leads to a set of new six new predictions, similarly to (13).

The same procedure is repeated by simulating the cases that the household signed up in August, September, ..., up to November. The predicted month is now earlier than the sign-up month, as in Section III-C1 (cfr. (12)).

Furthermore, a prediction for July can also be computed by considering that the household signed up in December, by applying the procedure described in Section III-C3, similarly to (14).

So, in total, $7 + 6 + 5 + 4 + 3 + 2 + 1 = 28$ individual

predictions are generated to compute the prediction

$$\hat{l}_7^{x_j} = \text{median}\left(\begin{array}{c} \frac{c^\star_{6,7}}{c^\star_{6,6}} \cdot l_6^{x_j}, \ldots, \frac{c^\star_{6,7}}{c^\star_{6,12}} \cdot l_{12}^{x_j}\Big|_{s=6} \\ \frac{c^\star_{7,7}}{c^\star_{7,7}} \cdot l_7^{x_j}, \ldots, \frac{c^\star_{7,7}}{c^\star_{7,12}} \cdot l_{12}^{x_j}\Big|_{s=7} \\ \frac{c^\star_{7,7}}{c^\star_{7,8}} \cdot l_8^{x_j}, \ldots, \frac{c^\star_{7,7}}{c^\star_{7,12}} \cdot l_{12}^{x_j}\Big|_{s=8} \\ \vdots \\ \frac{c^\star_{7,7}}{c^\star_{7,11}} \cdot l_{11}^{x_j}, \frac{c^\star_{7,7}}{c^\star_{7,12}} \cdot l_{12}^{x_j}\Big|_{s=11} \\ \text{median}\left(l_7^{x_1^{7,j}}, \ldots, l_7^{x_{n_{\text{similar}}}^{7,j}}\right)\Big|_{s=12} \end{array}\right). \quad (15)$$

Each row in the parentheses considers a different sign-up month from June ($s = 6$) up to December ($s = 12$). This is also illustrated in Fig. 1c.

In general, for a household x_j that signed up in month s, this data augmentation step creates extra predictions so that we have a total of $\sum_{i=1}^{13-s} i$ individual predictions that are used to compute the median.

E. Post-processing: smoothing the predictions

As a final step, the predictions are smoothed using a standard moving average technique. The motivation behind doing so is to reduce the effect of weather patterns in the year of the given measurements (see Section II for details). By using a moving average the weather effects are mitigated as the final prediction is the average between nearby months. This step has a parameter w describing the window length for the moving average. We tested three values as part of the IEEE-CIS competition and observed that, although the window size plays little role, five months is marginally better.

IV. Case Study

A. Problem statement IEEE-CIS Competition

The problem tackled in this paper is the monthly electrical consumption forecasting of thousands of households one year ahead using the historical consumption of the previous year. The proposed method was one of the award-winning methods in the IEEE-CIS competition [3]. For the sake of open-access and reproducibility, we use in this paper the same case study as it allows any other researcher to verify the results and re-use the data. In the competition, the goal was to forecast the monthly consumption of 3248 households in 2018 using (i) half hourly smart meter data from 2017 for the same households, with a different data availability for each household, ranging from one to twelve months representing different customer sign-up months, (ii) the weather data from 2017 for the households location at a daily resolution and (iii) additional household-specific information such as the number of occupants, the number of bedrooms, etc.

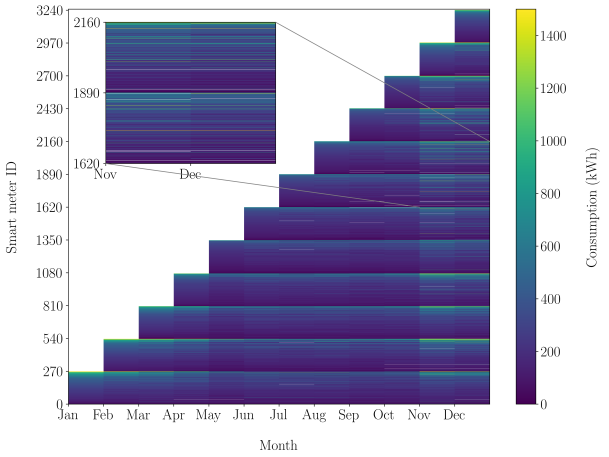


Fig. 4. Heat map of the monthly electrical consumption per smart household (in kWh). The y-axis represents the individual profiles and the x-axis the month. The color gives an indication of the monthly values. In the colder months (November & December) the consumption is higher. The faint white lines are missing data. It is clear from the staircase structure that only the first 270 profiles have twelve months of data available. The next 270 profiles have eleven months of data available, and so on up to the last 270 profiles which have only the month December.

B. Dataset

The dataset is provided by E.ON UK plc., in the context of the IEEE-CIS competition on energy prediction from smart meter data [3]. It consists of half hourly sampled time series describing the electrical consumption of 3248 households during the year 2017. It is a real world dataset containing all the challenges of a real application.

First, the profiles have a **different range of available historical data**, acknowledging that customers might have joined the measurement campaign at different times during the year, as is illustrated in Fig. 4. It is a two-dimensional representation of the electrical consumption recorded by the 3248 smart meters. It can be seen that the first 270 profiles (about a 12^{th} of the profiles) have twelve months of historical data available (from January until December), the next 270 profiles have eleven months of historical data (from February until December) and so on until the last part of the profiles which have only one month of historical data available (December).

Secondly, **there are missing days within the available months**, which are the faint white lines in Fig. 4.

Thirdly, additional household-specific information were collected through surveys, such as the type of building, the number of rooms, the number of occupants, and so on. However **these data are very sparse**, for example there are 48% of missing data in the type of building, 42% in the number of bedrooms, and always more than 97% for all the other additional information.

Finally, the competition also provides time series weather data, e.g., average, maximum and minimum daily temperature, associated with each household, in 2017. However, **the weather data of 2018 is not made available**, as well as the household consumption data of 2018. There is thus no validation dataset. It is subsequently not possible to assess the performance between the years. The performance could be checked on a leaderboard via the organizers. Further details

about the data are provided on the competition website [3].

The objective of the competition was to predict the monthly consumption (in kWh) of the 3248 households for the year 2018 as well as the aggregated yearly consumption for the same year. As explained by the organizers, the yearly total consumption, computed as the sum of the monthly forecasts, is of interest for billing purposes while the monthly values are used by the energy trading teams to buy the right amount of electricity on the energy market. As the data are real and publicly available, we use it here for our case study. As explained in the next section, we also consider the metrics proposed by the organizers of the competition.

C. Performance Metrics

We use the metrics from the competition to measure the performance of our method [3]. It is the weighted average error of the yearly prediction error and the monthly prediction error. In particular, the yearly relative absolute error (year_{rAE}) is measured as

$$\text{year}_{\text{rAE}} = \frac{\frac{1}{n}\sum_{j=1}^{n}|\hat{l}^{x_j} - l^{x_j}|}{\frac{1}{n}\sum_{j=1}^{n}|l^{x_j} - \bar{l}|} \tag{16}$$

where n is the total number of households, $\hat{l}^{x_j} = \sum_{i=1}^{12}\hat{l}_i^{x_j}$ is the predicted total yearly consumption of household x_j, l^{x_j} is the true total yearly consumption of household x_j and $\bar{l} = \frac{1}{n}\sum_{j=1}^{n}|l^{x_j}|$. Then the relative monthly error ($\text{month}_{\text{rAE}}$) is considered

$$\text{month}_{\text{rAE}} = \frac{1}{n}\sum_{j=1}^{n}\frac{\frac{1}{12}\sum_{i=1}^{12}|\hat{l}_i^{x_j} - l_i^{x_j}|}{\frac{1}{12}\sum_{i=1}^{12}|l_i^{x_j} - \bar{l}^{x_j}|} \tag{17}$$

where n is the total number of households, $\hat{\mathbf{L}}^{x_j} = [\hat{l}_1^{x_j}, \ldots, \hat{l}_{12}^{x_j}]$ is the predicted monthly consumption of household x_j , $\mathbf{L}^{x_j} = [l_1^{x_j}, \ldots, l_{12}^{x_j}]$ is the true monthly consumption of household x_j and $\bar{l}^{x_j} = \frac{1}{12}\sum_{i=1}^{12}|l_i^{x_j}|$.

Finally, both metrics are considered equally important and aggregated as

$$\text{total}_{\text{rAE}} = \frac{1}{2}\text{month}_{\text{rAE}} + \frac{1}{2}\text{year}_{\text{rAE}}. \tag{18}$$

D. Benchmark Models

As part of the competition, we tested several different prediction models. In this paper, we compare our model, with a baseline model that we designed and two well-known models that were submitted to the competition by another contestant: a linear regression (LR) model and an Autoregressive Integrated Moving Average (ARIMA) model [27]. These have been submitted by Dr. Kasun Bandara [28], [29]. In addition, we outline the performance of the top three approaches (one of which is the one proposed in this paper).

The naive benchmark we proposed, computes the monthly average over the available months of 2017 for each individual household and uses this monthly average as a prediction for each of the months of the year 2018.

Wenlong Wu (first place) [3], [30] proposed a Machine Learning pipeline consisting of data preprocessing, feature engineering, algorithm modeling, post-processing, and ensemble fusion. Fuzzy C-Means is applied, extracting twelve clusters of profiles. The model is based on a Light Gradient Boosting Machine (Light GBM) which makes use of bagging and boosting techniques. First, this model is applied on all the households individually, it learns the overall trend and makes day-level energy prediction. Secondly, one model is trained per cluster. Thirdly, a prediction is computed using the mean of November and December values. These three predictions are fused using a weighted average. Additionally, Wenlong Wu found 32 profiles with zero consumption values for November and December, which might be due to drained batteries during the acquisition of training data. Assuming that this problem is present in the collection of the ground-truth (i.e., test) data as well, he manually set their predictions to zero. We do not find this "battery trick" operation fair in terms of evaluation, because it exploits a specific problem in data collection. Finally a scaling step is applied, down-scaling the summer months and up-scaling the winter months as he noticed this improved the performance. The weather data and additional household-specific information were not used. The six features used by Wenlong Wu are categorical household ID, mean meter reading of the household, and time-related features like day of week, day of month and month encoded cyclically using sine and cosine functions, such that January and December are defined as similar.

Steffen Limmer's approach (second place) [3] is based on k nearest neighbors, the following sequence of steps is considered: data preprocessing, outlier removal (by computing the base distributions and filtering out abnormal distributions using isolation forest), an ensemble strategy to predict the monthly consumption and finally scaling. In the prediction step, the distribution of a meter is predicted as the average over the distributions of k nearest base meters (without the outliers) and then the monthly consumption is computed based on the distribution prediction and the December consumption.

E. The proposed approach with the battery trick

In addition to our proposed scalable ensemble approach that ranked third in the competition, we also consider its modified version that uses the *battery trick*, which is applied in the method that ranked first and explained in Section IV-D. This modification would enable our proposed approach to outperform the first ranked method in the competition, as shown in Section IV-G.

F. Run times

The proposed method takes around 45 minutes to run on a machine with 9th Generation "Coffee Lake" 2.6 GHz 6-Core Intel Core i7 mobile processor (I7-9750H), MacOS operating system, Python version 3.8.3: 20 seconds to generate the clusters, 19 minutes to determine the closest cluster for all inference cases of all profiles, 25 minutes to build the predictions and 40 seconds for ensemble and smoothing. Besides the training and ensemble/post-processing pipelines,

the method is parallelised on 4 threads. It could easily be parallelised with more threads to speed up the process even more.

The Light GBM model, by Wenlong Wu, takes 1h45 to run on the same machine[1]. It should also be noted that this method requires one light GBM model per smart meter individually, which reduces its applicability and scalability.

The run time of the KNN-Isolation forest ensemble method, by Steffen Limmer, could not be measured as the code was not made available.

G. Results & discussion

The results are presented in Table I. The ratio based approach described in this paper is compared with several standard time series techniques presented in Section IV-D. These techniques have been considered as they are often used and recognized as accurate methods in forecasting problems. From Table I, three main conclusions can be drawn. First, it is clear that for each method the monthly performance is worse than the yearly performance. This makes sense as the stochasticity due to human behavior and the influence of external factors such as the weather, averages over the course of one year, the error is thus reduced. Secondly, the performances of the LR and ARIMA models are lower than the more complex and advanced top three methods. Finally, it can be noticed that the proposed method with the additional battery trick ranks first, with a higher scalability than the light GBM model.

Table II presents a comparison of the top three methods' performance metrics, when the faulty smart meters are removed from the evaluation. The removed faulty smart meters are the same 32 smart meters treated by the "battery trick" detailed in Section IV-E. The method proposed in this paper outperforms the two other approaches.

A *t*-test [31] was performed on the two sets of monthly errors to assess the statistical difference between the performance of Light GBM model (which ranked first in the competition) and the proposed scalable ensemble approach with the battery trick (which outperforms Light GBM). The null hypothesis is that there is no statistical difference between the performance metrics of both methods. The p value obtained is 0.00958. The null hypothesis is thus rejected and we can conclude that there is a statistical difference in performance metrics between both methods with with $p < 0.005$. The same *t*-test was performed on the two sets of monthly errors after removing the faulty smart meters. The p value obtained is 0.01233, the null hypothesis is thus again rejected and we can conclude that there is a statistical difference in performance metrics between both methods with with $p < 0.005$.

H. Hyperparameter sensitivity study of the proposed method

The method makes use of only three hyperparameters through the different steps: the minimum cluster size $n_{\text{min,cluster}}$, the number of nearest neighbors n_{similar}, and the window length w for smoothing. For the following parameter values, we have observed insignificant difference in the total absolute error: $w = 2, 3, \cdots, 7$; $n_{\text{similar}} = 10, 20, \cdots, 100$; $n_{\text{min,cluster}} = 5, 6, \cdots, 20$. The best performance was obtained with the minimum cluster size $n_{\text{min,cluster}} = 10$, the number of nearest neighbors $n_{\text{similar}} = 50$, and the window length $w = 5$. See Appendix B for additional results. The method is robust to the values of the hyperparameters, as independently independently of the chosen hyperparemeters, the total relative absolute error remains in the interval $[0.6860, 0.7243]$.

Method	Monthly rAE	Yearly rAE	Total rAE
Our proposed approach (Original submission modified to include the battery trick)	0.9802	0.2861	0.6332
1st place: Light GBM (Original submission includes the battery trick)	1.0078	0.2864	0.6471
KNN-Isolation forest ensemble (Original submission modified to include the battery trick)	1.06905	0.28633	0.67769
2nd place: KNN-Isolation forest ensemble (Original submission includes outlier detection)	1.0728	0.2875	0.6801
3rd place: Our proposed approach (Original submission, does not include the battery trick)	1.0828	0.2892	0.6860
ARIMA	1.3852	0.3844	0.8848
Linear Regression	1.4461	0.3333	0.8897
Naive baseline (30th place)	1.4947	0.4345	0.9646

TABLE I. A comparison of three benchmark methods (naive baseline, linear regression and ARIMA) and the top three methods in the IEEE-CIS contest [3]. The methods proposed in this paper, are highlighted. Total rAE stands for relative absolute error, which is the weighted average error of the yearly prediction error and the monthly prediction error.

Method	Monthly rAE	Yearly rAE	Total rAE
Our proposed approach	0.9757	0.2885	0.6321
Light GBM	1.0165	0.2816	0.6491
KNN-Isolation forest ensemble	1.0654	0.2887	0.6771

TABLE II. A comparison of the top three methods' original submissions in the IEEE-CIS contest when removing the faulty smart meter out of the evaluation [3]. The method proposed in this paper, is highlighted. Total rAE stands for relative absolute error, which is the weighted average error of the yearly prediction error and the monthly prediction error.

V. Conclusion

In this paper a method has been proposed to predict the monthly and yearly electrical consumption of individual households one year ahead based solely on historical data of the previous year. The approach consists of different steps applied sequentially: pre-processing, data augmentation and normalization, clustering, prediction based on ratios and finally

[1]We have executed the code that was provided by the authors on the same machine.

ensemble learning and post processing. This novel method fills a gap in the literature by combining a forecast horizon (one year ahead) and forecast granularity (individual household) which has not been tackled yet. The model offers three main advantages. (i) Low data requirement: the method works even for households with only one month of historical data and it does not require additional household-specific information nor weather data. (ii) Scalability: the method has low computational requirements and does not involve household-specific information collection through questionnaires. (iii) Accuracy: the use case is based on data from the IEEE-CIS competition and the third position in the competition demonstrates the high performance of the model. The battery trick places the method in the first position, without altering the scalability. However, we think it is not a fair modification because it aims to estimate problems in the ground-truth data to improve the metrics, which wouldn't provide any benefit in practice.

In future work, the method could be improved by detecting outlier household profiles and defaults (i.e., measurement errors) in the historical data and treating these with a customized model. The model presented in this paper could also be applied on new datasets in order to assess the generalization of the method. As suggested in the literature [9], the approach could also be applied in the context of higher granularity forecasting by aggregating household predictions to the national level.

ACKNOWLEDGMENT

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program, from the KU Leuven Research Fund (projects C16/15/059, C3/19/053, C24/18/022, C3/20/117, C3I-21-00316) and from the European Research Council under the European Union's Horizon 2020 research and innovation programme (ERC Adv. Grant grant agreement N^{o}885682).

The work of this paper has been carried out as part of a collaboration between VITO/EnergyVille and the research groups STADIUS and DTAI from KU Leuven. The authors greatly thank the other contestants: Wenlong Wu, Steffen Limmer and Kasun Bandara for providing useful information and kindly answering our questions. The authors would like to thank Prof. Isaac Triguero for setting up the competition and E.ON UK plc. for providing the data.

REFERENCES

[1] N. Uribe-Pérez, L. Hernández, D. de la Vega, and I. Angulo, "State of the Art and Trends Review of Smart Metering in Electricity Grids," *Applied Sciences*, vol. 6, no. 3, pp. 1–24, 2016.

[2] T. Ahmad, H. Zhang, and B. Yan, "A review on renewable energy and electricity requirement forecasting models for smart grid and buildings," *Sustainable Cities and Society*, vol. 55, no. October 2019, p. 102052, 2020.

[3] I. Triguero, "IEEE-CIS Technical Challenge on Energy Prediction from Smart Meter Data," 2020. [Online]. Available: https://ieee-dataport.org/competitions/ieee-cis-technical-challenge-energy-prediction-smart-meter-data

[4] S. Ben Taieb and R. J. Hyndman, "A gradient boosting approach to the Kaggle load forecasting competition," *International Journal of Forecasting*, vol. 30, no. 2, pp. 382–394, 2014.

[5] X. Shao, C. S. Kim, and P. Sontakke, "Accurate deep model for electricity consumption forecasting using multi-channel and multi-scale feature fusion CNN-LSTM," *Energies*, vol. 13, no. 8, p. 1881, 2020.

[6] R. E. Abdel-Aal; and A. Z. Al-Garni, "Forecasting monthly electric energy consumption in eastern Saudi Arabia using univariate time-series analysis," *Energy*, vol. 22, no. I, pp. 1059–1069, 1997.

[7] C. W. Fu and T. T. Nguyen, "Models for Long-Term Energy Forecasting," in *2003 IEEE Power Engineering Society General Meeting, Conference Proceedings*, 2003, pp. 235–239.

[8] N. Cetinkaya, "Long-term electrical load forecasting based on economic and demographic data for Turkey," in *CINTI 2013 - 14th IEEE International Symposium on Computational Intelligence and Informatics, Proceedings*. IEEE, 2013, pp. 219–223.

[9] K. B. Lindberg, P. Seljom, H. Madsen, D. Fischer, and M. Korpås, "Long-term electricity load forecasting: Current and future trends," *Utilities Policy*, vol. 58, no. May, pp. 102–119, 2019.

[10] X. M. Zhang, K. Grolinger, M. A. Capretz, and L. Seewald, "Forecasting Residential Energy Consumption: Single Household Perspective," in *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*. IEEE, 2019, pp. 110–117.

[11] T. Y. Kim and S. B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," *Energy*, vol. 182, pp. 72–81, 2019.

[12] N. Fumo and M. A. Rafe Biswas, "Regression analysis for prediction of residential energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 332–343, 2015.

[13] R. Yao and K. Steemers, "A method of formulating energy load profile for domestic buildings in the UK," *Energy and Buildings*, vol. 37, no. 6, pp. 663–671, 2005.

[14] S. Haben, J. Ward, D. Vukadinovic Greetham, C. Singleton, and P. Grindrod, "A new error measure for forecasts of household-level, high resolution electrical energy consumption," *International Journal of Forecasting*, vol. 30, no. 2, pp. 246–256, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.ijforecast.2013.08.002

[15] M. R. Braun, H. Altan, and S. B. Beck, "Using regression analysis to predict the future energy consumption of a supermarket in the UK," *Applied Energy*, vol. 130, pp. 305–313, 2014.

[16] Y. Ma, J. Q. Yu, C. Y. Yang, and L. Wang, "Study on power energy consumption model for large-scale public building," in *Proceedings - 2010 2nd International Workshop on Intelligent Systems and Applications, ISA 2010*. IEEE, 2010, pp. 10–13.

[17] A. Rice, S. Hay, and D. Ryder-Cook, "A limited-data model of building energy consumption," in *BuildSys'10 - Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, 2010, pp. 67–72.

[18] J. F. M. Pessanha and N. Leon, "Forecasting long-term electricity demand in the residential sector," in *Procedia Computer Science*, vol. 55. Elsevier Masson SAS, 2015, pp. 529–538.

[19] H. X. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586–3592, 2012.

[20] M. Parti and C. Parti, "Total and Appliance-Specific Conditional Demand for Electricity in the Household Sector." *Bell Journal of Economics*, vol. 11, no. 1, pp. 309–321, 1980.

[21] T. Olofsson and S. Andersson, "Long-term energy demand predictions based on short-term measured data," *Energy and Buildings*, vol. 33, no. 2, pp. 85–91, 2001.

[22] S. Z. Selim and M. A. Ismail, "K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 1, pp. 81–87, 1984.

[23] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[24] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[25] Z. H. Zhou, *Ensemble methods: Foundations and algorithms*. CRC Press, 2012.

[26] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[27] J. Lago, F. De Ridder, and B. De Schutter, "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms," *Applied Energy*, vol. 221, no. November 2017, pp. 386–405, 2018.

[28] K. Bandara, R. Godahewa, and H. Hewamalage, "IEEE CIS Technical Challenge 2020 Git Hub Repository," 2020. [Online]. Available: https://github.com/rakshitha123/IEEE_CIS_Comp

[29] ——, "A Scalable Ensemble of Global and Local Models for Long-term Energy Demand Forecasting," in *41st International Symposium on Forecasting*, 2021. [Online]. Available: https://www.linkedin.com/posts/bandarakasun_isf2021-activity-6815901412514439168--u3s

[30] W. Wu, "IEEE CIS Technical Challenge 2020 Git Hub Repository," 2020. [Online]. Available: https://github.com/waylongo/cis-challenge-energy-prediction

[31] Student, "The Probable Error of a Mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.

APPENDIX A
DETAILS ON THE RESULTS

To better understand the approach proposed in this paper, it can be divided into several subcomponents and the impact of each one of them on the final performance metric can be analyzed. We start from the naive baseline detailed in Section IV-D, with a $\text{total}_{\text{rAE}}$ of 0.96. Then a prediction was built based on the sign-up month. That lead to a more accurate prediction, reducing the $\text{total}_{\text{rAE}}$ to 0.86.

The second step considered was ensemble learning and data augmentation. The same average relative consumption across all meters was used but assuming that a meter could have signed up in different months. This improved the accuracy of the prediction down to a $\text{total}_{\text{rAE}}$ of 0.80.

Third, the clustering based on *k*-means was introduced for meters that signed up in November or earlier, which improved the prediction to 0.75 $\text{total}_{\text{rAE}}$.

Fourth, k-nearest neighbors was considered for meters that signed up in December, which improved it to 0.72 $\text{total}_{\text{rAE}}$.

Fifth, post-processing and smoothing were added, which improved the prediction to a $\text{total}_{\text{rAE}}$ of 0.69.

Finally, 32 profiles with zero values for November and December were detected and their prediction was set to zero. This improved the prediction to a $\text{total}_{\text{rAE}}$ of 0.63.

In short, each of the individual components of the method has a similar effect as each yields a reduction between 0.03-0.05 in the $\text{total}_{\text{rAE}}$.

APPENDIX B
DETAILS ON THE HYPERPARAMETER SENSITIVITY STUDY

Independently of the chosen hyperparemeters, the total relative absolute error remains in the interval $[0.6801, 0.7243]$, as is shown in Table III. The mean and standard deviation of the total relative absolute error are respectively equal to $0,6991$ and $0,01034$.

Min cluster size	Nb of nearest neighb.	Window size	Monthly rAE	Yearly rAE	Total rAE
10	50	2	1.1039	0.2860	0.6950
10	50	3	1.0993	0.2857	0.6925
10	50	4	1.1039	0.2860	0.6950
10	50	5	1.0728	0.2875	0.6801
10	50	6	1.1039	0.2860	0.6950
10	50	7	1.0987	0.2913	0.6950
10	10	5	1.0918	0.2879	0.6898
10	20	5	1.0908	0.2876	0.6892
10	30	5	1.0908	0.2873	0.6890
10	40	5	1.0909	0.2871	0.6890
10	60	5	1.0909	0.2869	0.6889
10	70	5	1.0912	0.2868	0.6890
10	80	5	1.0917	0.2868	0.6892
10	90	5	1.0919	0.2868	0.6893
10	100	5	1.0922	0.2867	0.6895
10	50	5	1.1208	0.2922	0.7065
5	50	5	1.1269	0.3146	0.7208
6	50	5	1.1459	0.3026	0.7243
7	50	5	1.1358	0.2994	0.7176
8	50	5	1.1328	0.2987	0.7157
9	50	5	1.1012	0.2898	0.6955
10	50	5	1.1193	0.2905	0.7049
11	50	5	1.1141	0.2901	0.7021
12	50	5	1.1254	0.2896	0.7075
13	50	5	1.1210	0.2903	0.7056
14	50	5	1.1257	0.2920	0.7088
15	50	5	1.1212	0.2911	0.7061
16	50	5	1.1175	0.2904	0.7040
17	50	5	1.0973	0.2875	0.6924
18	50	5	1.0972	0.2879	0.6926
19	50	5	1.1181	0.2890	0.7035
20	50	5	1.1074	0.2886	0.6980

TABLE III. Comparison of the performance metrics of the method when varying the values of the three hyperparameters. The chosen hyperparameters are highlighted.