A data-driven Household Electricity Synthesiser for South Africa using Enveloped Sum of Gaussians

M.J. Ritchie [a], J.A.A. Engelbrecht [a], M.J. Booysen [a,*]

[a] *Department of E&E Engineering, Stellenbosch University, South Africa*

Abstract

Accurate modelling of household electricity on a large scale is a cornerstone for demand management and decarbonisation efforts. This paper introduces a data-driven model for South African households. According to the Paris Agreement, a 50% reduction in global CO_2 emissions is required to ensure the global average temperature does not surpass 2°C above pre-industrial levels. In this context, 26% of the overall European energy consumption is used by the residential sector and the global energy consumption is estimated to increase by 1.3% on average per year until 2050. In the last decade, the development and implementation of smart grid technologies have grown to efficiently and cost-effectively meet the electricity demands of the grid and mitigating greenhouse gas emissions. We present a data-driven, enveloped sum of Gaussians-based model and residential household synthesiser that statistically models the household's electricity usage demand based on smart meter data and generates synthetic data that accurately represents the actual demand profile, load peaks and daily variances for an individual or aggregate group of households. The measured data was gathered over a one year period for 1,200 households in South Africa. Our model accounts for temporal variations such as seasonality and the day of week, household uniqueness and is fully autonomous. Results showed that the root mean square error between the aggregated measured and synthetic electricity profiles were 0.181 A (5.68%) and the total energy of each profile were 75.9 and 76.3 A.h, respectively.

Keywords: Residential electricity demand modelling; Household electricity demand modelling; Demand-side management; Probabilistic modelling

1. Introduction

In the United States, 32% of primary energy consumption was contributed by the building sector in 2019, an 11% increase from that recorded in 2010 and global energy consumption is estimated to increase by 1.3% on average per year until 2050 [1, 2]. In Europe, 26% of the overall energy consumption is used by the residential sector [3]. In this context, a global CO_2 emission reduction of at least 50% by 2050 is required to guarantee the global average temperature does not surpass 2°C above pre-industrial levels [4, 5].

In the last decade, the development and implementation of smart grid technologies has significantly grown to efficiently and cost-effectively meet the electricity demands of the grid and mitigating greenhouse gas emissions [6, 7]. Smart grids enable the implementation of demand-side management strategies to help energy providers reshape demand profiles and reduce peak loads [8]. The management and planning of smart grids and buildings require that the electricity demand is accurately characterised [6]. Various building energy modelling approaches in the past commonly worked with monthly, seasonal and annual average demand profiles where the individual household's level of consumption is understated [9]. However, household smart meters have introduced new objectives for residential load modelling and forecasting by providing new important information of the occupant consumption patterns to improve the individual or aggregate household forecast accuracy [10, 11, 9]. These electricity consumption patterns have also shown to vary on a daily, weekly and seasonal basis [5]. Other factors include calender information (i.e. holidays or special events) and weather variables [12, 9, 13].

*Corresponding author
Email address: mjbooysen@sun.ac.za (M.J. Booysen)

For a smart grid strategy to be effective, vast amounts of real data over long time periods are required. The ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) hosted a Great Energy Predictor III competition which provided a publicly-available dataset, available at https://www.kaggle.com/c/ashrae-energy-prediction, that is commonly used as benchmark data for comparing the performance of various models [14, 15]. The data was collected from 1448 building from 16 various sources located in North America and Europe [16].

Although the smart meters collection of large amounts of time-of-use data with high resolution has enhanced our understanding of user behaviour patterns, it has also shown that residential load demand is still highly unpredictable due to the stochastic nature of the occupants to use devices at different times and day to day lifestyle variations [17]. Many studies that model residential demand work with datasets produced at high sampling rates (1 second - 15 minutes), however, low sampling rate datasets (15 minutes - 1 hour) are naturally more deterministic and allow for simpler methods to be used, such as probability functions and sum of Gaussians [18, 13].

There exist few studies that conduct research on representative residential load profiles for developing counties where low consumption households are typically treated as outliers and removed from the dataset [19, 20, 21]. South Africa is a developing country that experiences a lack of electricity generation capacity to meet the demand of the grid [22]. This has forced Eskom, the state-owned electricity supplier of the country, to incorporate rolling blackouts since 2008 as a result of lacking research and knowledge of the general electricity and energy demand [23, 24].

From early research, residential load profile modelling is divided into two groups: bottom-up and top-down models. Bottom-up load profile modelling determines the individual household's electricity consumption of each household appliance, occupant behaviour patterns and their usage relation with each appliance to obtain the household's electricity usage profile. On the other hand, top-down models determine the household's electricity consumption by using stochastic predictors or macro-variables such as the occupant family type and their behaviour, weather conditions and historically measured consumption data [25, 26]. More recently, a new category has been introduced that combines elements of the other two - the hybrid model [9, 13].

In this paper, we develop an enveloped sum of Gaussians (ESOG)-based household synthesiser that accurately captures the patterns of daily electricity usage profiles on an individual household level, and allows for reconstructing synthetic profiles that represent the demand characterisation, peaks and variance of the actual energy usage of a single or aggregate group of households. The next section reviews various methods proposed in literature that belong to these aforementioned models, and following this, we present by the novel contributions of our top-down approach.

2. Literature review and contributions

2.1. Household electricity models

Various household electricity models existing in the literature are reviewed in this section and are categorised based on the type of method: technical standard, clustering, stochastic, data-learning, time-series and data-driven. A recent review of the current state of these data-learning and time-series studies showed that 19% of them focused on residential buildings, 67% used real data, 31% used datasets longer than a year, 16% performed medium to long-term predictions, and 47% focused on overall energy demand [14]. Table 1 summarises all of the models that we reviewed and their corresponding key factors.

Technical standard models

These models represent a standard load profile (SLP) that is typically a large-scale representation of residential and/or commercial energy demand that are available for design purposes. A common disadvantage of these models is that they assume that there is not a strong variation in the residential load usage patterns for different households [58, 13].

An example of a technical standard model is the H0 SLP that is used in Germany and Austria to represent the countries' residential energy demand. The model uses a sampling rate of 15 minutes and produces profiles for seasons and differentiates between workdays and weekends [13].

Table 1: Summary of models and work that model residential load profiles. The symbols in the Approach column indicate if the work used a bottom-up (B), top-down (T) or Hybrid (H) approach. An S in the Day column indicates that days were differentiated only by workdays and weekends and a tick indicates that the differentiation extended to the day of the week. A dash indicates that the source didn't specify the information. The "Long." column indicates if the model performs long-term forecasting, the "Real" column indicates if real data was used, the "Dev." column indicates if the data was sourced from a developing country, and the "Unique" column indicates if the model accounted for household individuality.

Ref	Approach	Category	Residential	Long.	Real	Dev.	Season	Day	Unique	Res.
[27]	T	Technical standard	✓	✗	✗	✓	✓	S	✗	15 min
[28]	T	Clustering	✓	✗	✗	✓	✗	✗	✗	-
[29]	T	Clustering	✓	✓	✓	✓	✓	S	✗	1 hr
[30]	T	Clustering	✓	-	✓	✓	✗	✗	✗	-
[31]	T	Clustering	✓	✓	✓	✓	✗	✗	✗	1 hr
[32]	B	Stochastic	✓	-	-	✗	✗	✗	✗	15 min
[33]	B	Stochastic	✓	-	✗	✗	-	-	-	15 min
[34]	B	Stochastic	✓	✗	✗	✗	✓	✗	-	1 hr
[25]	B	Stochastic	✓	-	✗	✗	✓	S	✗	1 hr
[35]	T	Stochastic	✓	-	✗	✗	✓	S	✗	1 hr
[36]	B	Stochastic	✓	✗	✗	✗	✓	S	✓	1 min
[18]	T	Stochastic	✓	-	✗	✗	✗	✗	✓	30 min
[37]	H	Stochastic	✓	✗	-	-	✓	✗	✗	1 min
[38]	B	Stochastic	✓	✓	✗	✓	✓	-	✗	1 hr
[39]	T	Stochastic	✓	-	✓	✗	✗	✗	-	1 min
[12]	B	Stochastic	✓	✓	✗	✗	✓	S	-	10 min
[40]	T	Stochastic	✓	-	✗	-	✗	S	✗	1 hr
[41]	B	Stochastic	✓	-	✗	✗	✓	S	✗	1 hr
[42]	H	Stochastic	✓	-	✗	✗	✓	-	-	1 s
[43]	B	Stochastic	✓	-	✓	✗	✓	✓	✗	10 min
[44]	B	Stochastic	✓	-	✗	✗	✓	S	✓	5 min, 1 hr
[45]	H	Stochastic	✓	-	✗	✗	✓	S	✗	1 min
[46]	T	Stochastic	✓	-	✓	✓	-	✓	-	2 s
[47]	T	Data-learning	✓	✗	✓	-	-	✗	✗	-
[48]	T	Data-learning	✓	✓	✓	✓	✗	✗	-	-
[49]	T	Data-learning	✗	✗	✓	✗	✓	S	✓	1 hr
[50]	T	Data-learning	✓	✗	✓	✗	✓	S	✓	15 min
[51]	T	Data-learning	✗	✗	✓	✗	-	✓	-	1 hr
[52]	T	Data-learning	✗	✓	✓	✗	-	-	-	-
[53]	T	Data-learning	✗	✗	✗	✓	✓	S	-	1 hr
[54]	T	Data-learning	✓	✗	✓	✓	✓	S	-	-
[55]	T	Time-series	✗	✗	✗	✗	✓	✓	✓	1 hr
[56]	T	Time-series	✗	✗	✓	✗	✓	✓	✓	-
[57]	T	Data-driven	✓	✓	✓	✓	✗	✗	✗	-
This paper	T	Data-driven	✓	✓	✓	✓	✓	✓	✓	1 hr

Clustering models

These models are based on approaches that group a set of households together based on their electrical demand profile similarities. Despite the name referring to clustering techniques, we also include models that use non-clustering techniques.

Zhou et al. [30] developed a fuzzy clustering technique for electricity consumption patterns mining from household data collected from 1200 households situated in Jiangsu Province, China. The results of this model classified 938 valid households into four and six groups. By extracting the characteristics of each group, it was proven that the electricity consumption patterns of each household was effectively clustered.

Heunis and Dekenah [29] presented a model that estimates the household electricity consumption of South African households by specifying sociodemographic variables such as household income, appliance ownership/usage and occupant employment/education. This is used to produce a load profile that is represented as a typical, or average, profile with standard deviations of the probable usage at each hour of the day.

Clustering models are disadvantageous due to their generalisation of household consumption patterns and disregarding the usage that is strongly dependent on the individual household.

Stochastic models

These models use probabilistic distributions, cumulative probability functions or conditional demand analysis to model the electrical demand profiles [13]. Many of these models make use of Monte Carlo and

Markov chain statistical techniques [32, 33, 36, 18, 37, 12, 40, 42].

Ortiz et al. [41] proposed a bottom-up stochastic method with emphasis on modelling the variations between occupants and their stochastic usage of various electrical appliances, whereas many other approaches base electrical consumption on statistical data. Their model was implemented in a TRNSYS (Transient System Simulation Tool) environment and a Spanish apartment building was simulated, and the output of their model is random realistic energy consumption profiles. Since there is a relationship between the model inputs and the labelling of electrical appliances, their model can also determine potential energy savings from high-performance appliances. A disadvantage of this model, and the majority of bottom-up models, is that it is very dependent on the input data and high modelling intensity.

Labeeuw and Deconinck [40] present a top-down model of representing residential electricity loads for large-scale smart grid strategies. However, acquiring large databases of detailed residential load profiles can be a difficult task due to privacy laws. They overcome this problem by using statistical data and Monte Carlo simulations for 1300 residential load profiles. Similar profiles are grouped together using Mixture Model clustering. For each cluster, two Markov models are created to generate a load profile based on a randomised sequence of states: the first model produces a behaviour model of the cluster, and the second model randomises the behaviour. Their results showed that 86.4% of their profiles fitted the original data well, and 13.6% of the profiles, which represented extreme cases of high electricity demand, underestimated the original data by half or a third.

Data-learning models

These models develop a forecasting model based on machine learning methods, such as neural networks. Edwards et al. [50] predicted the future hourly residential energy consumption of residential buildings using several machine learning methods to compare their performance. Their case study used sensor data collected from three residential households and their results showed that LV-SVM (least squares support vector machines) was the best method to predict residential household demand. Additionally, they performed their case study using the ASHRAE Greatest Energy Prediction dataset to determine the performance of the methods on commercial buildings. Interestingly, they found that LV-SVM performed the worst of all the methods, and a neural network-based method had the best performance.

Time-series models

These models forecast a system's future behaviour by analysing historical data to make statistical conclusions based on time correlation. The original data is decomposed into trend, seasonality and residual components. More information on time-series forecasting can be found in [59].

Penya et al. [55] evaluated the performance of the most popular short-term load forecasting models in literature: an ARIMA (autoregressive integrated moving average) time-series model, a neural network, an AR (autoregressive) time-series model, and a Bayesian network. They concluded that the AR model obtained the best performance, where the average MAPE (mean absolute percentage error) was 14.3 for three day-ahead forecastings, and increased to 18.4 for six day-ahead forecastings. However, their study used a small sample size and the buildings that were measured were non-residential.

Data-driven models

These models predict the residential demand by analysing the relationships and patterns of actual data without using matching usage to electrical devices and not using data-learning or time-series methods.

Xu et al. [57] present a clustering-based probability distribution model that simulates electricity usage in residential households. They used smart meter data for 86 672 households located in the Jiangsu Province, China. The measured data is clustered using a two-step k-means clustering methods. The first step extracts features that represent the level of electricity usage, and the second step clusters each of the clusters to extract usage patterns. Following this, probability distributions are fitted to the data within each sub-cluster. The outcome of this approach produced 16 sub-clusters and results suggested the model had precise prediction accuracy and adequate parameter estimation. Although the model only determined the monthly electricity demand, it was stated that similar results could also be produced for shorter time frames if data with finer time resolution is used.

Key factors missing in existing models

Although many bottom-up approaches exist, top-down models are advantageous over bottom-down models as they don't require information of every appliance as well as having a lower level of complexity since they do not need to model the usage of every appliance [13]. Many of the studies do not model the profile for residential households, account for all temporal variations as well as household uniqueness. There also exists few studies applicable for developing countries.

Although the surge of machine learning and increasing accessibility to data has shown good opportunities for data-learning and time-series models, most of these studies in literature use non-residential datasets and produce short-term predictions of the load profiles. Many of these techniques used large datasets that extend back many years, whereas electricity consumption behaviour is expected to change within the next few years and decade due to increased usage (due to electronic devices, working from home, electro-mobility etc.) [13]. Furthermore, these models are also highly complex and require intense computational power.

2.2. Contributions

We propose a novel data-driven model that is based on the sum of Gaussian technique, namely enveloped sum of Gaussians, and clustering and statistical analyses to model an individual residential household's electrical usage profile and the aggregated energy demand of a group of households. A novel household synthesiser is also presented that generates realistic usage profiles from the developed model for the application of medium to long-term forecasting. Our model addressed the limitations determined in the literature review by differentiating usage patterns by season and the day of the week, has low computational complexity and is fully autonomous. We evaluate the accuracy of the model by comparing the synthetically generated profiles with measured data. The measured data was gathered over a one year period for 1,200 household smart meters in South Africa, a developing country. The measured data, source code for the ESOG-based model and household synthesiser, and synthetic profiles can be found in the supplementary material.

3. Household electricity usage model and synthesiser

In this section, we discuss the development of the household electricity usage synthesiser that models the day-to-day electricity usage behaviour of a household based on historically measured electrical meter data, and synthesiser new days by generating the daily peaks from statistical models. The ESOG-based household electricity usage model is described in Section 3.2, followed by the synthesiser in Section 3.3.

3.1. Electricity usage data

The electricity usage data used in this paper was obtained from the Domestic Electrical Load Metering Data database, conducted at the University of Cape Town [60]. The database contains a large volume of hourly meter profiles for many households (1994-2014), where each data measurement represents the aggregated current (Amps) used over that hour. We use usage profiles for 1200 residential households, each spanning one year, in South Africa. Since a household operates with a constant voltage, we use the terms *energy* and *electricity* in the remainder of this paper to refer to current-based usage profiles as energy is implied.

3.2. ESOG-based model

Our method aims to describe the measured usage profiles for each household using a data-driven model. The model is then used to generate synthetic usage profiles that match the original profiles (shapes and quantum of energy) statistically.

The steps used in our method are shown in Figure 1 and detailed below. First, we fit a curve to describe the usage profile of each day for each household: Gaussians are used to describe the amplitude, width, and time of each of the peaks a day's profile. The envelope of the Gaussians is used to describe the the dynamic variable component (DVC) of the day's profile. This is different to the typically used sum of Gaussians, since we only include the largest Gaussian for an hour into the profile for that hour.

For each household, these fitted profiles are used to cluster over days according to the Gaussians' peak amplitude, peak width, and time of occurrence. For each of the clusters a probability of occurrence for any given day is calculated for weekdays and weekend days, and for each season. These clusters and their

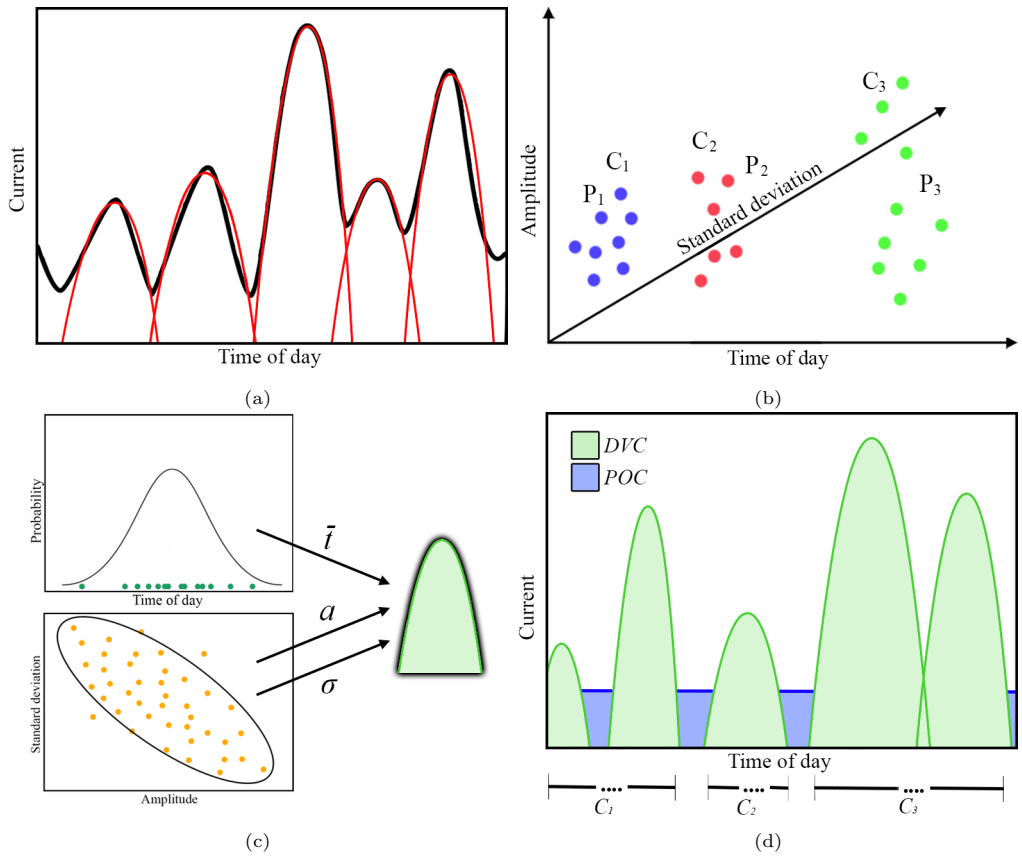


Figure 1: a) Gaussians (red) are fitted to all of the peaks identified in the usage profile (black) for a single day. b) The time of day, amplitude and standard deviation of the fitted Gaussians for all of the days are clustered in terms of the time of day to produce clusters C_1, C_2 and C_3 with their corresponding probability of occurrence P_1, P_2 and P_3. c) A probabilistic distribution is fitted to the time-of-day component and a joint probabilistic distribution is fitted to the amplitude and standard deviation components of all the data points belonging to a cluster. These distributions are used to generate a synthetic peak. d) A synthetic profile is constructed from two components: 1) the dynamic variable component (DVC) which are the synthetically generated peaks occurring throughout the day, and 2) the proportional offset component (POC) which is the supplementary energy filled between these peaks.

probabilities are then used to construct the DVC of a synthetic usage profiles for a household. Since this DVC only describes the energy around the peaks, a proportional offset component (POC), akin to a non-linear base load profile, is added to the enveloped profile to ensure the day's aggregate energy is also representative. This POC is constructed by scaling the household's mean seasonal profile, such that the total energy for the synthetic day matches that of a measured day with similar energy in the DVC. The POC is treated like any of the Gaussians to determine the enveloped profile

The model accounts for variations due to seasonality by separating the data into four seasonal datasets and deriving a model for each one. The model also accounts for user behavioural variations due to the day of the week by separating each seasonal dataset into seven datasets that correspond to each day of the week and deriving a model for each one. Therefore, a total of 28 datasets and models will be established for a household to ensure that the usage patterns are effectively differentiated.

3.2.1. Dynamic variable component

Given the usage profile for a single day, the statistical properties of the usage behaviour can be derived by fitting Gaussians to all of the peaks to capture their time of occurrence, amplitude and energy. Doing so will ensure that most of the energy in the profile is mathematically measured due to its oscillating nature. A peak finding algorithm (using the SciPy signal processing Python library) is used to identify all the peaks in the profile. For each peak, a Gaussian is fitted, using

$$f(x, \bar{t}, a, \sigma) = a \cdot e^{-(\frac{x-\bar{t}}{\sigma})} \quad (1)$$

where the mean $\bar{t}$ and amplitude a is set to that of the identified peak and the standard deviation σ of each Gaussian is chosen so that the total energy within the Gaussian is equivalent to that of the identified peak. The process of matching Gaussians to the identified peaks is shown in Figure 1a.

The matched Gaussians for all of the days are superimposed onto a single 24-hour cycle to identify repeating patterns in the actual peaks. Each Gaussian is replaced with a single data point that represents its time of day, amplitude and width. The peak data is grouped into clusters by using k-means clustering [61] to cluster the data in terms of the time of day, as shown by the 3-dimensional plot in Figure 1b. The optimal number of clusters is determined by using the elbow method [62]. The clusters are indicated by C_i, where subscript i is the index of the cluster, and each cluster represents a repeating usage pattern. Associated with each cluster is the probability of occurrence, indicated by $P_{i,j}$, which represents the probability of j peaks occurring for cluster i on a given day, where $j = 0, 1, 2..., n$ and $\sum_{j=1}^{n} P_{i,j} = 1$. For simplicity, we use P_i to refer to all of the probabilities of occurrence for cluster i.

Next, we determine the probabilistic distribution of the time of day of the repeating peaks in a cluster by fitting a Gaussian probability density function to the time-of-day component of all the data points belonging to that cluster, as shown by the top graph in Figure 1c. For cluster i, the Gaussian probability density function is determined by calculating the time-of-day mean $\bar{t}_i$ and variance $\sigma^2_{t,i}$ for all the data points in the cluster. We also determine the probabilistic distribution of the peak amplitudes and standard deviations in a cluster. Since the amplitudes and standard deviations of the peak may be statistically dependent, we fit a Gaussian joint probability density function to amplitude and standard deviation components of all the data points belonging to a cluster, as shown by the bottom graph in Figure 1c. For cluster i, the Gaussian joint probability density function is determined by calculating the amplitude and standard deviation means $\bar{a}_i$ and $\bar{\sigma}_i$, variances $\sigma^2_{a,i}$ and $\sigma^2_{\sigma,i}$, and the correlation coefficient ρ_i. The probabilistic distributions for the time-of-day, amplitude and standard deviation components are used to generate a *synthetic peak* with mean $\bar{t}$, amplitude a and standard deviation σ, as shown by Figure 1c.

3.2.2. Proportional offset component

The previous section describes how the peaks, or DVC, of an electricity usage profile can be statistically modelled and used to generate synthetic peaks. It is also important to model the remaining energy in the electricity usage profile, or POC, to ensure that all of the energy in the profile is captured by our model. This is determined after all the peaks have been matched with Gaussians. The POC for day d is calculated as follows:

$$E_{\mathrm{POC,d}} = E_{\mathrm{total,d}} - E_{\mathrm{DVC,d}} \quad (2)$$

where $E_{\mathrm{total,d}}$ and $E_{\mathrm{DVC,d}}$ are the total and DVC energy for the profile of day d and the DVC energy is calculated by summing the envelope of all of the Gaussian distributions for that day.

We determine a probabilistic distribution of the total energy and DVC energy for all of the days by fitting a Gaussian joint probability density function to these components. This is performed in a similar way to the joint density function determined for the peak amplitudes and standard deviations shown in Figure 1c. The Gaussian joint probability density function is determined by calculating the total energy and AC energy means $\bar{E}_{total}$ and $\bar{E}_{DVC}$, variances σ^2_{total} and σ^2_{DVC}, and the correlation coefficient ρ_E. Given the DVC energy $E_{\mathrm{DVC,d}}$ for a daily profile, the POC energy $E_{\mathrm{POC,d}}$ can be determined by obtaining the total energy $E_{\mathrm{total,d}}$ from the Gaussian joint probability density function and using Equation 2.

3.3. Electricity usage synthesiser

Algorithm 1 shows the procedure used to generate a synthesised electricity usage profile from a household's ESOG-based model.

If we assume that we want to generate an electricity usage profile for a certain number of days, we iterate through all of the days. For each day, we first iterate through all of the clusters and generate random synthetic peaks for each cluster. To generate a synthetic peak, we determine the time of the peak by drawing a random sample from the time of day Gaussian distribution and the amplitude and standard deviation of the peak by

drawing a random sample from the Gaussian joint distribution for the given cluster. Given the time of peak $\bar{t}$, amplitude a and standard deviation σ, we construct the synthetic peak by using the following equation:

$$I(t) = a \cdot e^{-(\frac{t-\bar{t}}{\sigma})} \quad \forall\, t \tag{3}$$

where I(t) is the instantaneous current as a function of time. This shows that the synthetic peak is a Gaussian. The synthetic profile for the day is constructed by superimposing all of the synthetic peaks. Next, we add the POC to the profile to ensure that the overall energy of the profile represents a realistic amount of energy for the household on that day. To do this, we first calculate the amount of DVC energy in the current synthetic usage profile using the following equation:

$$E_{DVC} = \sum_{t=t_i}^{t_f} I(t) \tag{4}$$

where t_i and t_f are the first and last time instants for the day. We then determine the desired amount of POC energy E_{POC} for the profile by using E_{DVC}, the Gaussian joint distribution discussed in the previous section to determine E_{total}, and using Equation 2. This ensures that the chosen amount of POC energy is equivalent to an actual day that had a similar amount of DVC energy. The next step involves iteratively increasing the energy at each time instant of the profile in small increments until $E_{DVC} + E_{POC} = E_{total}$. We desire that the POC energy is first added to the time instants where the energy is the lowest. We achieve this by setting the POC energy offset height h to zero, then iteratively increasing the offset height by a very small value h_{inc}, and at each iteration we ensure that the energy at all time instants is above or equal to h. An example of how a synthetic day is constructed from the DVC and POC energy is shown in Figure 1d. We have assumed that the offset height h remains constant throughout the day, however, we will further alter

Algorithm 1 Electricity usage synthesiser algorithm

procedure GENERATE SYNTHESISED USAGE DATA
  for `each day` **do**
    DVC
    for `each cluster i` **do**
      $x \leftarrow$ *Roll number between 0 and 1* // where x is a random variable
      $m \leftarrow 0$
      while $m < n+1$ *and* $x > \sum_{k=1}^{m} P_{i,m}$ **do**
        $m \leftarrow m+1$ // where m is the number of peaks generated for cluster i
      end while
      for `each generated peak j` **do**
        $\bar{t}_{i,j} \leftarrow \mathcal{N}(\bar{t}_i, \sigma^2_{t,i})$ // generate one sample
        $a_{i,j}, \sigma_{i,j} \leftarrow \mathcal{N}(\bar{a}_i, \bar{\sigma}_i, \sigma^2_{a,i}, \sigma^2_{\sigma,i}, \rho_i)$ // generate one sample
        $I_j(t) \leftarrow a_{i,j} \cdot e^{-(\frac{t-\bar{t}_{i,j}}{\sigma_{i,j}})} \quad \forall\, t$
      end for
    end for
    Superimpose all generated peaks to obtain synthesised DVC energy profile $I(t)$
    $E_{DVC} \leftarrow \sum_{t=t_i}^{t_f} I(t)$ // determine the DVC energy for generated profile

    POC
    $E_{total} \leftarrow \mathcal{N}(\bar{E}_{total}, \bar{E}_{DVC}, \sigma^2_{total}, \sigma^2_{DVC}, \rho_E, E_{DVC})$ // generate one sample
    $E_{POC} \leftarrow E_{total} - E_{DVC}$
    $h \leftarrow 0$ // where h is the current POC height
    while $\sum_{t=t_i}^{t_f} I(t) - E_{DVC} < E_{POC}$ **do**
      $h \leftarrow h + h_{inc}$ // where h_{inc} is a very small value
      for `each time instant t` **do** // iterate through each instant for the day
        if $I(t) < h$ **then**
          $I(t) \leftarrow h$
        end if
      end for
    end while
  end for
end procedure

the shape of the offset component so that h at each time instant of the day is proportional to the mean daily profile determined for the household for that season. This ensures that the supplementary energy is added appropriately to the profile according to the household's general trend of usage.

Figure 2a shows an example of fitting Gaussians (red) to the actual data (blue). Figures 2a-c compare the envelope of all of the fitted Gaussians (orange) to the actual data (blue) for a) a good fit (0.0 A RMSE, 0.0 % RMSPE), b) an average fit (0.12 A RMSE, 6.28 % RMSPE), and c) a bad fit (0.39 A RMSE, 44.5 % RMSPE). These figures show how nearly all of the profile is captured by the envelope of fitted Gaussians, and portions not captured are contributed to the POC energy.

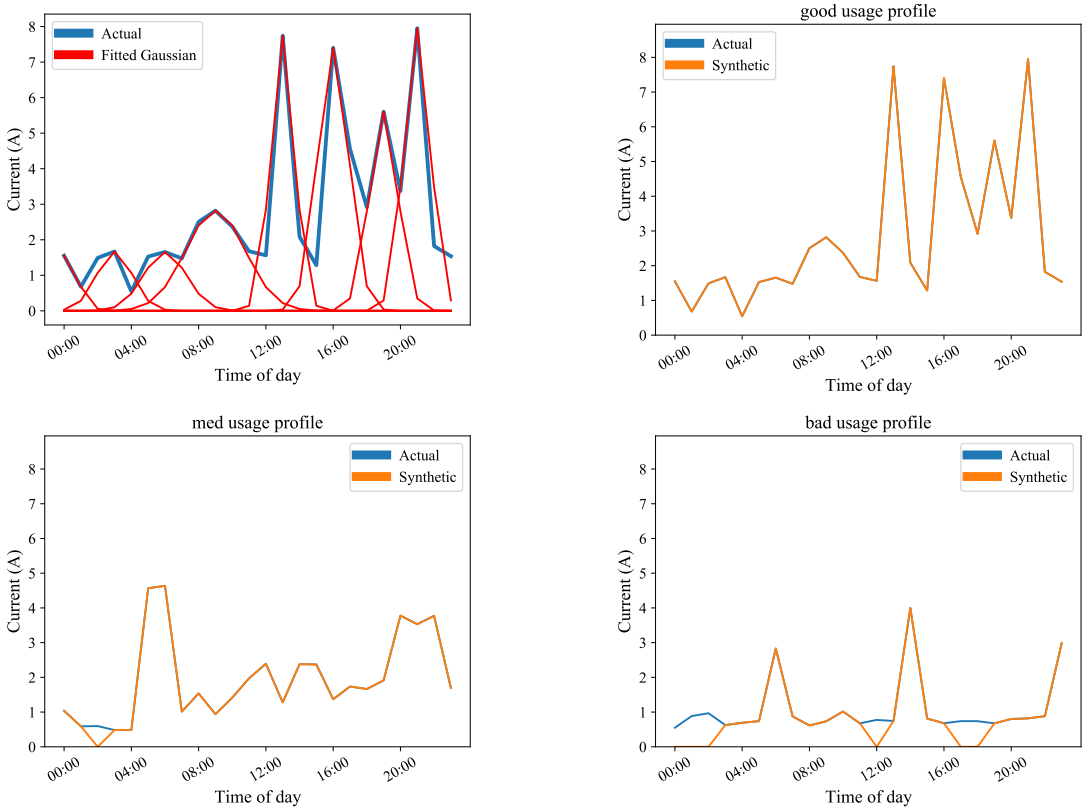


Figure 2: a) Example of fitting Gaussians (red) to the actual data (blue). b-c) Comparison of the envelope of all of the fitted Gaussians (orange) to the actual data (blue) for a) a good fit (RMSE: 0.0 A, RMSPE: 0.0 %), b) an average fit (RMSE: 0.12 A, RMSPE: 6.28 %), and c) a bad fit (RMSE: 0.39 A, RMSPE: 44.5 %).

4. Results

The results are obtained by fitting our model to one year (12 weeks per season) of actual electricity usage data for 1,200 households and, similarly, one year of synthetic electricity usage data is generated for each household. We determine the performance of the ESOG-based model by assessing the following metrics:

1. How well does our model characterise the actual data? (Section 4.1)
2. How well does our model capture the variations of a household's daily energy usage profile (Section 4.2)
3. Does our model accurately represent the aggregated load of the grid? (Section 4.3)

4.1. Descriptive matching

In this section, we determine how well our model describes the actual data by determining how closely-matched the envelope of Gaussian distributions, or fitted profiles, is to the actual daily profile and how well the most significant peak for each day is modelled. The DC energy component is included with the fitted profiles to ensure that the total energy is matched.

4.1.1. Profile matching

Figures 3a and 3b show box plots of the RMSE and RMSPE between the actual daily profiles and the corresponding fitted profiles for all seasons, for only summer and only winter. The results are shown for the bad, average, and good cases which correspond to the RMSE and RMSPE of daily profiles that correspond to the 95^{th}, 50^{th} and 5^{th} percentiles for each household. For all seasons, the distribution of errors, given as [25^{th} *percentile*, **median**, 75^{th} *percentile*], for the bad days of each household is [*0.0*, **0.086**, *0.153*] A ([*0.0*, **4.27**, *12.1*] %), [*0.0*, **0.010**, *0.055*] A ([*0.0*, **0.32**, *3.98*] %) for average days, and [*0.0*, **0.0**, *0.008*] A ([*0.0*, **0.0**, *0.35*] %) for good days. The results are almost identical for all the seasons and the only difference is observed for the marginally bigger errors in winter. As seen in Figure 2, these errors are mostly contributed by the DC portions of actual daily profiles not exactly matching that of the fitted profiles. But if it is the DC that is wrong, jsut add the DVC to each profile (so the total energy matches) before doing this RSME test. REsults will look much better

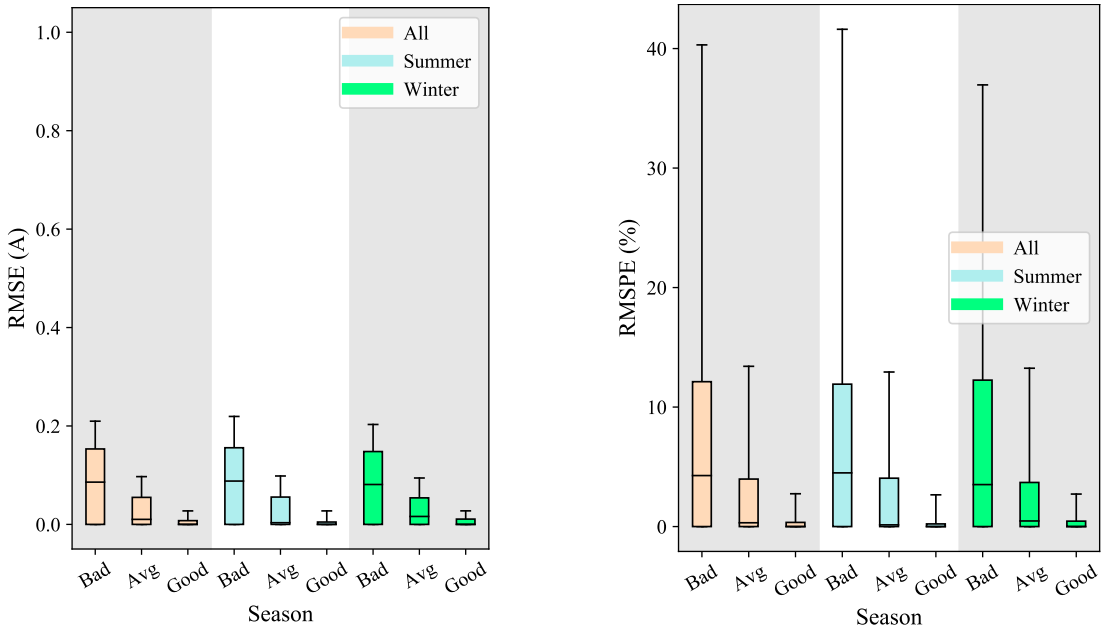


Figure 3: Box plots for a) the RMSE and b) RMSPE calculated between the daily actual and fitted profiles for the indicated seasons. The bad, average and good boxes refer to the RMSE and RMSPE of daily profiles corresponding to the 95^{th}, 50^{th} and 5^{th} percentiles for each household.

4.1.2. Largest peak matching

We determined the results of how well the fitted profiles modelled the largest peak of each actual daily profile in terms of amplitude and time of day for all households and season. There was no amplitude or time error observed for the comparison of the fitted and actual profiles for any of the days and this means that the largest peak was always perfectly modelled by the fitted profile.

4.2. Household characterisation

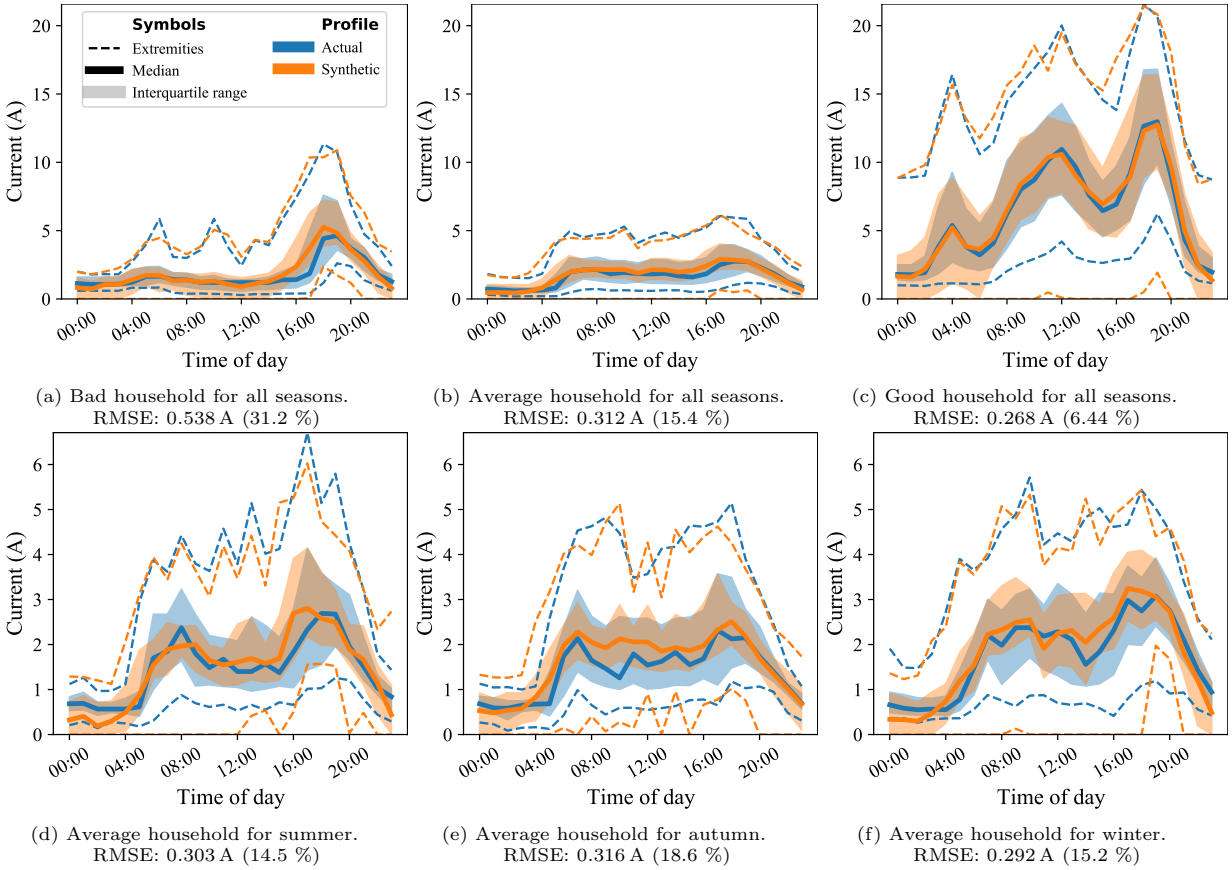
In this section, we determine how well our probabilistic characterisation of the variance of a household's synthetic energy usage profile describes the actual variations seen in that household's profile. Results are obtained for selected individual households of interest, as well as for all households.

4.2.1. Individual household usage

Figures 4 a-f compares the actual (blue) and synthetic (orange) usage profiles for a household, where the median (solid line), interquartile range (shaded area) and extremities (dashed line) of all the daily usage profiles occurring at each hour is shown. Figures 4a, 4b and 4c show the profiles for bad, average and good households, which indicates the households with the 95^{th}, 50^{th} and 5^{th} RMSE across all the days for all

seasons, respectively. Figures 4d, 4e and 4f show the profiles for the average household for summer, autumn and winter, respectively.

It can be seen that the that the median synthetic profile for the bad household overestimates the evening peak significantly. Despite the heavy usage of the good households, the synthetic profile succeeds to closely follow the actual usage profile and accurately represent the peaks. The RMSE for the median profile of the bad, average and good households is 0.538 A (31.2 %), 0.312 A (15.4 %) and 0.268 A (6.44 %). For the average household, the results are similar for all of the seasons, where the RMSE for the median profile for summer, autumn and winter is 0.303 A (14.5 %), 0.316 A (18.6 %) and 0.292 A (15.2 %). These figures also show that the upper extremity profile is well represented by the synthetic usage profiles, however, the lower extremity is not well represented. This occurs when a day is generated with few synthetic peaks, where the energy in these peaks already satisfies the total daily energy and negates the need for additional POC energy.



(a) Bad household for all seasons. RMSE: 0.538 A (31.2 %)

(b) Average household for all seasons. RMSE: 0.312 A (15.4 %)

(c) Good household for all seasons. RMSE: 0.268 A (6.44 %)

(d) Average household for summer. RMSE: 0.303 A (14.5 %)

(e) Average household for autumn. RMSE: 0.316 A (18.6 %)

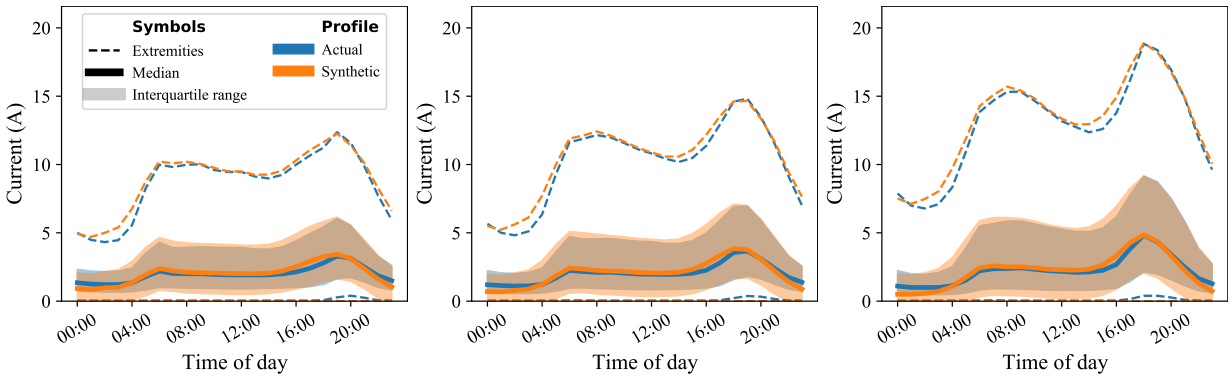
(f) Average household for winter. RMSE: 0.292 A (15.2 %)

Figure 4: Comparison of the actual (blue) and synthetic (orange) usage profiles for a household, where the median (solid line), interquartile range (shaded area) and extremities (dashed line) of all of the daily usage profiles occurring at each hour is shown. a-c) Profiles are shown for the a) bad, b) average and c) good households, where the 95^{th}, 50^{th} and 5^{th} RMSPE across all the days for all seasons. d-f) Profiles are shown for the average household for d) summer, e) autumn and f) winter. RMSE and RMSPE (in brackets) values are measured for the median profiles. I know we want to keep the same units, but we are only comparing d-f here. Should we make the units coser to 7A for d-f?

4.2.2. Overall household usage

Figures 5 a-f compares the actual (blue) and synthetic (orange) usage profiles for all households, where the median (solid line), interquartile range (shaded area) and extremities (dashed line) of all the daily usage profiles occurring at each hour is shown. Figures 5a, 5b and 5c shows the profiles for summer, all seasons, and winter and the RMSE for for the median profile for all of the households is 0.257 A (12.7 %), 0.282 A (13.7 %) and 0.314 A (14.0 %). For all of these figures, it can be seen that the evening peak is accurately represented by the median, interquartile range, and extreme case for all of the synthetic profiles. Furthermore, it can be seen that the RMSE is slightly larger for winter profiles. This is a result of more electricity used overall

during this season and, therefore, obtaining larger misprediction errors.



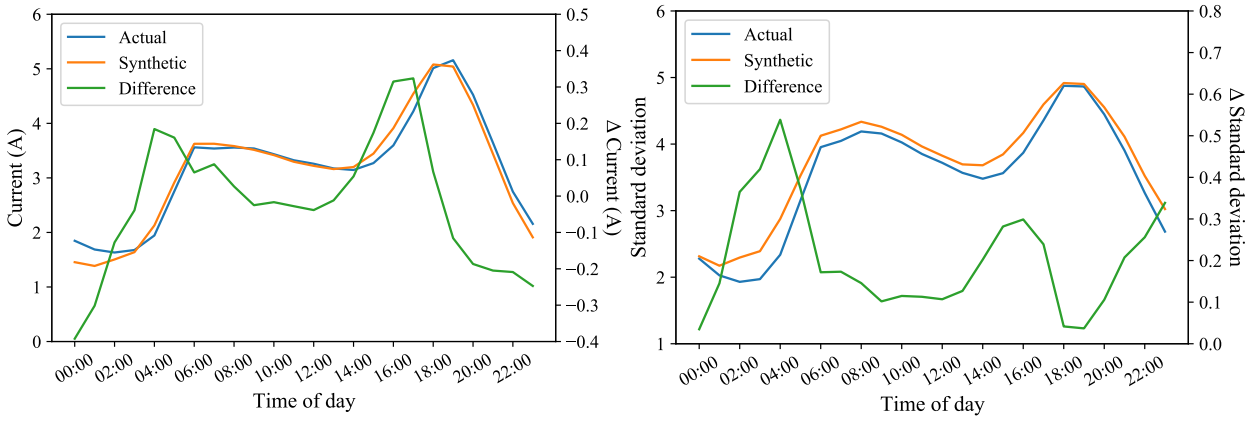
(a) All households for summer. RMSE: 0.257 A (12.7 %)

(b) All households for all seasons. RMSE: 0.282 A (13.7 %)

(c) All households for winter. RMSE: 0.314 A (14.0 %)

Figure 5: Comparison of the actual (blue) and synthetic (orange) usage profiles for all households in a) summer, b) all seasons and c) winter, where the median (solid line), interquartile range (shaded area) and extremities (dashed line) of all of the daily usage profiles occurring at each hour is shown. RMSE and RMSPE (in brackets) values are measured for the median profiles.

Table 2 shows the mean, standard deviation, and difference of the actual and synthetic median profile, measured across all households and seasons, for the each hour of the day. For example, the mean difference for the i^{th} hour is calculated as $\mu^i_{synthetic} - \mu^i_{actual}$ and the standard deviation difference is calculated as $\sigma^i_{synthetic} - \sigma^i_{actual}$. It can be seen that the mean of the synthetic profiles is larger than that of the actual profiles during peak hours and is smaller during off-peak hours. Despite these differences being negligibly small, it shows that our model possibly overestimates the energy in peaks and underestimates the POC energy, a result of how the augmentation of the POC energy profile is determined. Moreover, it can be seen that the standard deviation of the synthetic profiles is always larger than that of the actual profiles, where the difference is larger during peak hours and smaller during off-peak hours. This shows that the synthetic profiles generated from our model have more energy variation during each hour of the day to that of the actual profiles, where variations increase during peak hours.



(a)

(b)

Figure 6: To be written... Add 'A' to std figure... and point to 2ns y axis in legend...

Table 2: Table of the mean, standard deviation, and difference for the actual and synthetic median profile, measured across all households and seasons, for each hour of the day. The mean difference for the i^{th} hour is calculated as $\mu^i_{synthetic} - \mu^i_{actual}$ and the standard deviation difference is calculated as $\sigma^i_{synthetic} - \sigma^i_{actual}$.I think this should rather be a two plots than a table. Values on the primary y axis and Differences on the secondary y-axis

Hour	**Mean (μ)**			**Standard deviation (σ)**		
-	Synthetic	Actual	Difference	Synthetic	Actual	Difference
1	1.455	1.847	-0.392	2.314	2.279	0.035
2	1.386	1.687	-0.302	2.172	2.027	0.145
3	1.505	1.632	-0.127	2.295	1.93	0.365
4	1.638	1.678	-0.039	2.391	1.971	0.42
5	2.129	1.944	0.184	2.875	2.337	0.538
6	2.914	2.754	0.161	3.526	3.152	0.373
7	3.625	3.56	0.065	4.126	3.954	0.172
8	3.626	3.538	0.088	4.221	4.048	0.173
9	3.583	3.557	0.027	4.335	4.19	0.146
10	3.515	3.54	-0.025	4.261	4.159	0.102
11	3.418	3.434	-0.017	4.14	4.025	0.115
12	3.296	3.324	-0.028	3.964	3.851	0.113
13	3.221	3.259	-0.039	3.826	3.719	0.107
14	3.162	3.173	-0.012	3.695	3.568	0.127
15	3.199	3.145	0.054	3.682	3.479	0.202
16	3.445	3.271	0.174	3.845	3.563	0.282
17	3.911	3.596	0.315	4.167	3.869	0.299
18	4.543	4.219	0.324	4.595	4.356	0.239
19	5.079	5.013	0.067	4.917	4.876	0.042
20	5.042	5.158	-0.116	4.903	4.866	0.037
21	4.338	4.525	-0.187	4.553	4.447	0.106
22	3.44	3.645	-0.205	4.111	3.904	0.207
23	2.538	2.747	-0.209	3.522	3.266	0.256
24	1.91	2.157	-0.247	3.024	2.685	0.339
Overall	3.163	3.183	-0.02	3.954	3.756	0.197

4.3. Aggregated household usage

In this section, we determine if our model produces the same aggregate data, in terms of overall and peak energy, as the grid of our 1,200 households for a normalised 24-hour energy profile. We also compare the ability of our model to represent the aggregated profile with two other residential load models from the literature as well as a benchmark dataset.

Figure 7 shows a normalised 24-hour energy profile that compares the actual (blue) and synthetic (orange) grid daily energy usage for the mean (solid line), interquartile range (shaded area) and extremity (dashed line) cases. Our model's profile accurately follows the actual profile, except for the off-peak period in the early morning hours where the synthetic profile is lower than than the actual profile for the mean and lower quartile case, and higher for the upper extremity case. However, the synthetic data performs well at following the usage peaks which are of higher importance than the off-peak periods. The wider interquartile range for the synthetic demand also indicates that the synthetic profiles inherit a higher variety of energy demand throughout the day. The RMSE of our model and the actual mean demand is 0.181 A (5.68 % RMSE) and the overall energy demand is 75.9 A.h and 76.3 A.h, respectively. This confirms that our model excels at modelling the household overall and peak energy demand of the grid.

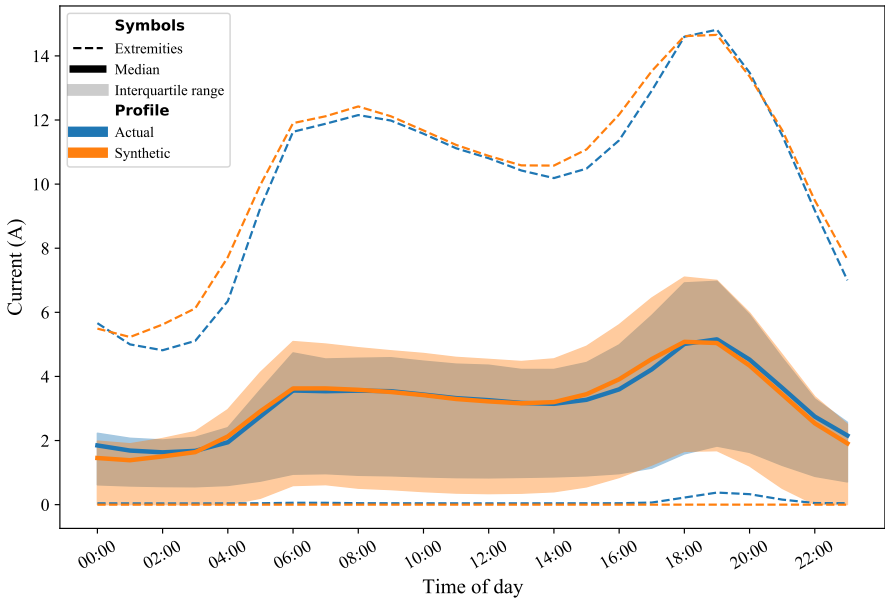


Figure 7: Comparison of the normalised actual (blue) and synthetic (orange) grid daily energy usage over 24 hours for 1200 households, where the mean (solid line), interquartile range (shaded area) and extremities (dashed line) is shown for all of the daily usage profiles. The mean profiles for the synthetic and actual usage data has a 0.181 A RMSE (5.68 % RMSPE), and the total energy usage for the synthetic and actual mean usage profiles is 75.9 A.h and 76.3 A.h.

We compare the results of our model with two other residential load profiles and one benchmark dataset: the Anvari model, a top-down stochastic approach developed by Anvari et al. [46] that uses German and Austrian data, the Heunis model, a clustering approached devised by Heunis and Dekenah [29] that uses South African data, and the ASHRAE load profile, a widely known benchmark dataset used in North America and Europe [15, 16]. Figure 8a shows the 24-hour normalised electricity demand for the actual, synthetic, Anvari, Heunis and ASHRAE data over the entire year, and Figure 8b shows the absolute relative error profiles for these profiles compared to the actual load profile. The Anvari, Heunis, and ASHRAE load profiles have been scaled to match the evening peak of the actual profile to provide a fair comparison of the profile shapes.

It has already been discussed in the previous figure that the synthetic profile accurately follows the actual demand. The Anvari profile shows that the the morning peak occurs one hour later in the day than that of the actual profile, and the Heunis profile overestimates the amplitude of the morning peak. Both of the Anvari and Heunis profiles notably underestimate the energy demand during the off-peak periods. The ASHRAE profile does not show a two-peak profile, but instead shows that there is high demand throughout the off-peak period between the morning and evening peaks. However, the ASHRAE dataset does not distinguish between energy consumption of residential households and lodges.

The RMSE and RMSPE of each profile when compared to the actual demand are 0.181 A (5.68%) for our model, 0.419 A (13.2%) for the Anvari model, 0.494 A (15.5%) for the Heunis model, and 1.164 A (36.6%) for the ASHRAE dataset.

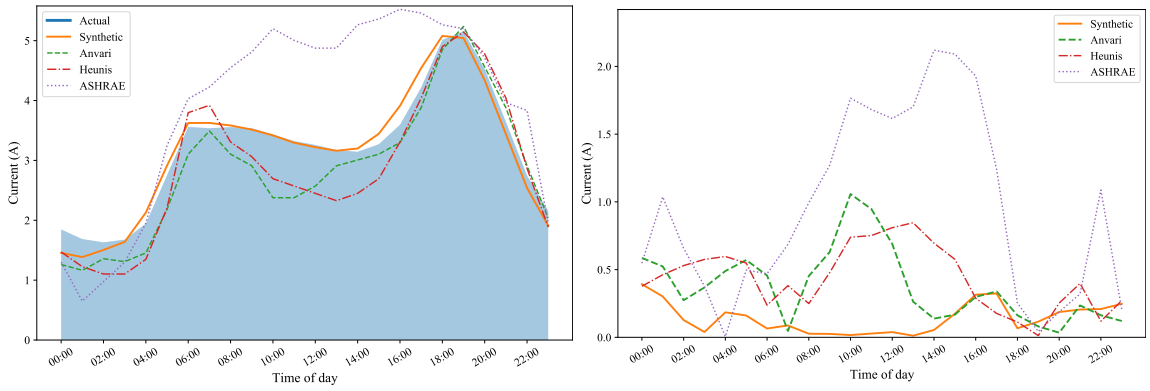


Figure 8: Comparison of the 24-hour normalised actual (shaded blue area), synthetic (orange solid line), Anvari (green dashed line), Heunis (red dash-dotted line) and ASHRAE (red dashed line) electricity demand for their a) average profiles and b) the absolute relative error profiles when compared to the actual profile.

Finally, Figure 9 shows a comparison of the daily average residential energy usage for various sources that use data from South Africa (This paper, Heunis and Dekenah [29]), other developing countries (Xu et al. [57], Yang et al. [31] and Zhou et al. [30]) and developed countries (Brecha et al. [63], Balaris et al. [64] and Yao and Steemers [34]). The figure shows that South African residential household usage is significantly low and the data from our paper shows that the country has a distribution ranging from almost zero to 34 kWh/day, which is due to the country's unique and extreme inequality amongst high and low consumption households. This also shows the importance of using South African data to synthesise realistic profiles that represent the country's energy demand.

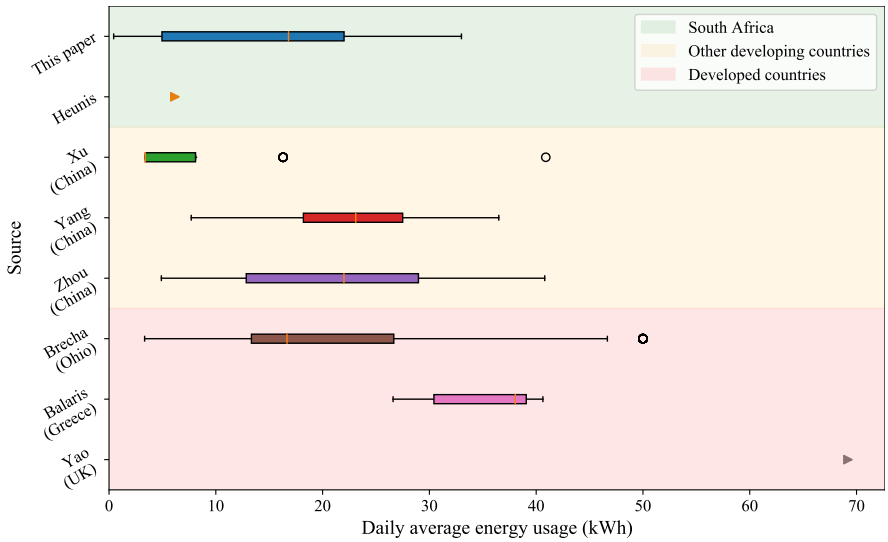


Figure 9: A comparison of the daily average residential energy usage for various sources that use data from South Africa, other developing countries and developed countries.

5. Conclusion

The ESOG-based model and household usage synthesiser that we develop in this paper statistically models a households electricity usage demand based on historically measured data. The model can accurately and autonomously model the residential power demand for individual household's and the aggregate demand for a sub-grid of household's, making it a useful tool for energy management systems. The results show that our model improves on existing household electricity models to characterise a household's electricity behaviour

that accounts for seasonality and day of week. A comparison of our synthetic aggregated usage data with the actual data showed that our model accurately represented the grid power demand with an RMSE and RMSPE of 0.181 A and 5.68 %, and the overall energy demand for the synthetic and actual demand profiles was 75.9 A.h and 76.3 A.h.

Acknowledgement

The authors thank Eskom and MTN South Africa for funding, and UCT, SANEDI and Wiebke Hutiri for making the data public.

References

[1] EIA, "Monthly energy review – June 2020. US Energy Information Administration," 2020.

[2] L. Zhang, J. Wen, Y. Li, J. Chen, Y. Ye, Y. Fu, and W. Livingood, "A review of machine learning in building load prediction," *Applied Energy*, vol. 285, p. 116452, 2021.

[3] European Commission and Directorate-General for Energy, *EU energy in figures : statistical pocketbook 2018*. Publications Office, 2018.

[4] S. UNFCCC, "Report of the conference of the parties on its twenty-first session, held in paris from 30 november to 13 december 2015. addendum. part two: Action taken by the conference of the parties at its twenty-first session," in *Bonn: United Nations Framework Convention on Climate Change*, 2015.

[5] D. Fischer, A. Surmann, and K. B. Lindberg, "Impact of emerging technologies on the electricity load profile of residential areas," *Energy and Buildings*, vol. 208, p. 109614, 2020.

[6] M. E. El-Hawary, "The smart grid—state-of-the-art and future trends," *Electric Power Components and Systems*, vol. 42, no. 3-4, pp. 239–250, 2014.

[7] A. Rahman, V. Srikumar, and A. D. Smith, "Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks," *Applied energy*, vol. 212, pp. 372–385, 2018.

[8] B. P. Esther and K. S. Kumar, "A survey on residential demand side management architecture, approaches, optimization models and methods," *Renewable and Sustainable Energy Reviews*, vol. 59, pp. 342–351, 2016.

[9] B. Yildiz, J. I. Bilbao, J. Dore, and A. B. Sproul, "Recent advances in the analysis of residential electricity consumption and applications of smart meter data," *Applied Energy*, vol. 208, pp. 402–427, 2017.

[10] P. G. Da Silva, D. Ilić, and S. Karnouskos, "The impact of smart grid prosumer grouping on forecasting accuracy and its benefits for local electricity market trading," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 402–410, 2013.

[11] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer, "Cluster-based aggregate forecasting for residential electricity demand using smart meter data," in *2015 IEEE international conference on Big data (Big data)*. IEEE, 2015, pp. 879–887.

[12] M. Muratori, M. C. Roberts, R. Sioshansi, V. Marano, and G. Rizzoni, "A highly resolved modeling technique to simulate residential power demand," *Applied Energy*, vol. 107, pp. 465–473, 2013.

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

[14] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, 2018.

[15] Vaishali Dhar, *ASHRAE- Great Energy Predictor III- A Machine Learning Case Study*, 2021, available at https://medium.com/analytics-vidhya/ashrae-great-energy-predictor-iii-a-machine-learning-case-study-a01a67eb048d.

[16] C. Miller, P. Arjunan, A. Kathirgamanathan, C. Fu, J. Roth, J. Y. Park, C. Balbach, K. Gowri, Z. Nagy, A. D. Fontanini *et al.*, "The ashrae great energy predictor iii competition: Overview and results," *Science and Technology for the Built Environment*, vol. 26, no. 10, pp. 1427–1447, 2020.

[17] Z. A. Khan and D. Jayaweera, "Approach for smart meter load profiling in monte carlo simulation applications," *IET Generation, Transmission & Distribution*, vol. 11, no. 7, pp. 1856–1864, 2017.

[18] F. McLoughlin, A. Duffy, and M. Conlon, "The generation of domestic electricity load profiles through markov chain modelling," *Euro-Asian Journal of Sustainable Energy Development Policy*, vol. 3, p. 12, 2010.

[19] H.-Â. Cao, C. Beckel, and T. Staake, "Are domestic load profiles stable over time? an attempt to identify target households for demand side management campaigns," in *IECON 2013-39th annual conference of the IEEE industrial electronics society*. IEEE, 2013, pp. 4733–4738.

[20] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 420–430, 2014.

[21] W. Toussaint and D. Moodley, "Comparison of clustering techniques for residential load profiles in south africa," 2019.

[22] R. Inglesi, "Aggregate electricity demand in south africa: Conditional forecasts to 2030," *Applied energy*, vol. 87, no. 1, pp. 197–204, 2010.

[23] Mail and Guardian, *Nersa: Power crisis cost SA about R50bn.*, 2008, available at https://mg.co.za/article/2008-08-26-nersa-power-crisis-cost-sa-about-r50bn/.

[24] A. Pouris, "Energy and fuels research in south african universities: a comparative assessment," 2008.

[25] J. V. Paatero and P. D. Lund, "A model for generating household electricity load profiles," *International journal of energy research*, vol. 30, no. 5, pp. 273–290, 2006.

[26] L. G. Swan and V. I. Ugursal, "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques," *Renewable and sustainable energy reviews*, vol. 13, no. 8, pp. 1819–1835, 2009.

[27] H0 SLP, *Standard Load Profiles (SLP)*, N/A, available at https://www.energienetze-bayern.com/de/strom/netzzugang/lastprofilverfahren/standardlastprofile--slp-.html.

[28] J.-T. Bernard, D. Bolduc, and N.-D. Yameogo, "A pseudo-panel data model of household electricity demand," *Resource and Energy Economics*, vol. 33, no. 1, pp. 315–325, 2011.

[29] S. Heunis and M. Dekenah, "A load profile prediction model for residential consumers in south africa," in *Twenty-Second Domestic Use of Energy*. IEEE, 2014, pp. 1–6.

[30] K. Zhou, S. Yang, and Z. Shao, "Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study," *Journal of cleaner production*, vol. 141, pp. 900–908, 2017.

[31] T. Yang, M. Ren, and K. Zhou, "Identifying household electricity consumption patterns: A case study of kunshan, china," *Renewable and Sustainable Energy Reviews*, vol. 91, pp. 861–868, 2018.

[32] C. F. Walker and J. L. Pokoski, "Residential load shape modelling based on customer behavior," *IEEE transactions on power apparatus and systems*, no. 7, pp. 1703–1711, 1985.

[33] A. Capasso, W. Grattieri, R. Lamedica, and A. Prudenzi, "A bottom-up approach to residential load modeling," *IEEE transactions on power systems*, vol. 9, no. 2, pp. 957–964, 1994.

[34] 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.

[35] J. Widén, M. Lundh, I. Vassileva, E. Dahlquist, K. Ellegård, and E. Wäckelgård, "Constructing load profiles for household electricity and hot water from time-use data—modelling approach and validation," *Energy and buildings*, vol. 41, no. 7, pp. 753–768, 2009.

[36] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: A high-resolution energy demand model," *Energy and buildings*, vol. 42, no. 10, pp. 1878–1887, 2010.

[37] O. Ardakanian, S. Keshav, and C. Rosenberg, "Markovian models for home electricity consumption," in *Proceedings of the 2nd ACM SIGCOMM workshop on Green networking*, 2011, pp. 31–36.

[38] Z. Ren, P. Paevere, and C. McNamara, "A local-community-level, physically-based model of end-use energy consumption by australian housing stock," *Energy policy*, vol. 49, pp. 586–596, 2012.

[39] C. Bucher and G. Andersson, "Generation of domestic load profiles-an adaptive top-down approach," in *Proceedings of PMAPS*, vol. 2012, 2012, pp. 10–14.

[40] W. Labeeuw and G. Deconinck, "Residential electrical load model based on mixture model clustering and markov models," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 3, pp. 1561–1569, 2013.

[41] J. Ortiz, F. Guarino, J. Salom, C. Corchero, and M. Cellura, "Stochastic model for electrical loads in mediterranean residential buildings: Validation and applications," *Energy and Buildings*, vol. 80, pp. 23–36, 2014.

[42] B. J. Johnson, M. R. Starke, O. A. Abdelaziz, R. K. Jackson, and L. M. Tolbert, "A matlab based occupant driven dynamic model for predicting residential power demand," in *2014 IEEE PES T&D Conference and Exposition*. IEEE, 2014, pp. 1–5.

[43] J. Munkhammar, J. Rydén, and J. Widén, "Characterizing probability density distributions for household electricity load profiles from high-resolution electricity use data," *Applied Energy*, vol. 135, pp. 382–390, 2014.

[44] A. Marszal-Pomianowska, P. Heiselberg, and O. K. Larsen, "Household electricity demand profiles–a high-resolution load model to facilitate modelling of energy flexible buildings," *Energy*, vol. 103, pp. 487–501, 2016.

[45] E. McKenna and M. Thomson, "High-resolution stochastic integrated thermal–electrical domestic demand model," *Applied Energy*, vol. 165, pp. 445–461, 2016.

[46] M. Anvari, E. Proedrou, B. Schäfer, C. Beck, H. Kantz, and M. Timme, "Data-driven load profiles and the dynamics of residential electric powerconsumption," *arXiv preprint arXiv:2009.09287*, 2020.

[47] F. Lai, F. Magoules, and F. Lherminier, "Vapnik's learning theory applied to energy consumption forecasts in residential buildings," *International Journal of Computer Mathematics*, vol. 85, no. 10, pp. 1563–1588, 2008.

[48] Q. Li, P. Ren, and Q. Meng, "Prediction model of annual energy consumption of residential buildings," in *2010 international conference on advances in energy engineering*. IEEE, 2010, pp. 223–226.

[49] D. M. Solomon, R. L. Winter, A. G. Boulanger, R. N. Anderson, and L. L. Wu, "Forecasting energy demand in large commercial buildings using support vector machine regression," 2011.

[50] R. E. Edwards, J. New, and L. E. Parker, "Predicting future hourly residential electrical consumption: A machine learning case study," *Energy and Buildings*, vol. 49, pp. 591–603, 2012.

[51] C. Roldán-Blay, G. Escrivá-Escrivá, C. Álvarez-Bel, C. Roldán-Porta, and J. Rodríguez-García, "Upgrade of an artificial neural network prediction method for electrical consumption forecasting using an hourly temperature curve model," *Energy and Buildings*, vol. 60, pp. 38–46, 2013.

[52] S. Ferlito, M. Atrigna, G. Graditi, S. De Vito, M. Salvato, A. Buonanno, and G. Di Francia, "Predictive models for building's energy consumption: An artificial neural network (ann) approach," in *2015 xviii aisem annual conference*. IEEE, 2015, pp. 1–4.

[53] D. Zhao, M. Zhong, X. Zhang, and X. Su, "Energy consumption predicting model of vrv (variable refrigerant volume) system in office buildings based on data mining," *Energy*, vol. 102, pp. 660–668, 2016.

[54] A. Yousaf, R. M. Asif, M. Shakir, A. U. Rehman, and M. S Adrees, "An improved residential electricity load forecasting using a machine-learning-based feature selection approach and a proposed integration strategy," *Sustainability*, vol. 13, no. 11, p. 6199, 2021.

[55] Y. K. Penya, C. E. Borges, D. Agote, and I. Fernández, "Short-term load forecasting in air-conditioned non-residential buildings," in *2011 IEEE International Symposium on Industrial Electronics*. IEEE, 2011, pp. 1359–1364.

[56] C. Fan, F. Xiao, and S. Wang, "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques," *Applied Energy*, vol. 127, pp. 1–10, 2014.

[57] J. Xu, X. Kang, Z. Chen, D. Yan, S. Guo, Y. Jin, T. Hao, and R. Jia, "Clustering-based probability distribution model for monthly residential building electricity consumption analysis," in *Building Simulation*, vol. 14, no. 1. Springer, 2021, pp. 149–164.

[58] D. Peters, R. Völker, F. Schuldt, and K. von Maydell, "Are standard load profiles suitable for modern electricity grid models?" in *2020 17th International Conference on the European Energy Market (EEM)*. IEEE, 2020, pp. 1–6.

[59] R. H. Shumway and D. S. Stoffer, "Time series analysis and its applications," *Studies In Informatics And Control*, vol. 9, no. 4, pp. 375–376, 2000.

[60] W. Toussaint, "Domestic electrical load metering, hourly data 1994-2014," Apr 2020. [Online]. Available: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/759

[61] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[62] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in k-means clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013. [Online]. Available: https://bit.ly/2VJsxOE

[63] R. J. Brecha, A. Mitchell, K. Hallinan, and K. Kissock, "Prioritizing investment in residential energy efficiency and renewable energy—a case study for the us midwest," *Energy Policy*, vol. 39, no. 5, pp. 2982–2992, 2011.

[64] C. A. Balaras, A. G. Gaglia, E. Georgopoulou, S. Mirasgedis, Y. Sarafidis, and D. P. Lalas, "European residential buildings and empirical assessment of the hellenic building stock, energy consumption, emissions and potential energy savings," *Building and environment*, vol. 42, no. 3, pp. 1298–1314, 2007.

Supplementary material

- The model and synthesiser source code, measured electricity usage data, and resulting synthetic profiles are available at https://bit.ly/SyntheticHotWater.