Generating synthetic residential electricity load profiles using household characteristics

Zahra Rahimpour, Navid Haghdadi, Mike Roberts, Nargess Nourbakhsh and Anna Bruce

School of PV and Renewable Energy Engineering and Collaboration on Energy and Environmental Markets, UNSW Sydney 2052

Introduction

Studying the impact of distributed energy resources (DER) and different operating strategies on outcomes for customers, aggregators and the electricity grid requires access to customer electricity load profiles. Moreover, demand-side management programs can be more customised when energy service providers know about household characteristics. However, many households to not have the metering required to collect and store interval consumption data, there are few public data sets of Australian residential load profiles; and given the diversity in household characteristics, these can not readily be used in lieu of real consumption data to assess DER outcomes for a specific customer and therefore to inform decisions about energy tariffs and rooftop solar and battery deployment. For this purpose, it would be useful to be able to estimate the expected shape of residential load profiles based on household and appliance characteristics.

Several past or current Australian projects have collected electricity consumption timeseries data of residential customers linked to household and appliance characteristics collected through household surveys. However only high-level statistics or benchmarks are published from these studies. One of these is the Energy Use Data Model (EUDM) dataset, collected by CSIRO as part of the National Energy Analytics Research (NEAR) program, which provides basic electricity consumption statistics for over 1800 Victorian households and over 600 Western Australian households, with analysis published in [1]. Another is the set of electricity consumption benchmarks for residential customers developed by Frontier Economics for the Australian Energy Regulator (AER). The benchmarks are typical annual and seasonal consumption figures for residential electricity usage in Queensland, New South Wales, the Australian Capital Territory, South Australia, Tasmania, and Victoria [2]. However, neither the raw data nor typical load profile *shape* for different household groups are published. Load profiles linked to surveys for about 4,000 households were published from the 2012-2013 Smart Grid, Smart City (SGSC) trial conducted by Ausgrid [3]. These profiles have been used by the research community for a range of analyses. The EUDM study used time series clustering to group electricity customers according to their load characteristics, however this method is not explicitly designed to assess the importance of customer characteristics that have an impact on load profile shape [4]. In another study the SGSC data was analysed to determine key drivers for residential peak demand on hot summer days [5]. However, to date, the SGSC data has not been used to assess the factors that are most important in determining the overall shape of the profiles. To date there have been no models published that can generate synthetic residential load profiles for customers with a diverse range of household characteristics.

This paper proposes a *normalised random forest* (NRF) model for generating synthetic customer load profiles based on household and appliance characteristics, using the SGSC dataset. The results demonstrate the NRF model outperforms the RF model in predicting load profiles and also its application to be used for important feature selection.

This paper progresses as follows. First, we briefly introduce the proposed RFN model. Next, in the results section, we compare the RF model with NRF, scaled with historical yearly data, and also compare the two models if feature selection methods such as permutation method and recursive feature elimination are applied. Finally, we conclude the paper with some directions for future work.

Methodology

In this section, first, we briefly review methods for estimating time series data and then describe the proposed NRF model. Next, we explain the two popular methods that can be used for calculating RF and NRF models' feature importance.

In the literature, many methods are used for forecasting time series data. The most common techniques are artificial neural networks (ANNs), support vector machines (SVMs), and autoregressive integrated moving averages (ARIMAs). *Random Forest* (RF) is a state-of-the-art ensemble-based supervised machine learning method which is less frequently used in the literature, but offers a promising solution compared to ANNs, SVMs, and ARIMAs [6-7]. In contrast to load estimation problems, which are types of regression problems, SVM is more suited to classification problems.. ARIMA has been used in applications when the historical data is used to predict the future time series. It is often less efficient in forecasting load patterns when other inputs such as categorical variables are also used.

RF models have been applied to a variety of machine learning tasks in recent years and can be trained on a combination of both categorical and numerical variables. RF work mainly on the basis of having multiple regression trees and averaging or voting the results produced by each regression tree. In order to produce the final output, the algorithm employs a process known as bagging. In more detail, an initial dataset is randomly sampled, and several dataset subsets are subsequently formed. Moreover, a set of features is randomly selected. The selection of random features significantly reduces the correlation between the datasets. Each of these subset datasets produces its own decision tree/learning model. As a result, the variance is greatly reduced in comparison to using a single decision tree. In the final step, the aggregation process is used to get the final result from the RF model. The new data for prediction is fed through all decision trees, each of which produces its own result. A vote or average of the results from all these trees determines the final outcome.

In this work, to generate a synthetic load profile, we developed and compared the performance of both RF and normalised random forest (NRF) models. The NRF is an RF model that uses normalised historical load data and customer characteristics as inputs to the model (refer to Figure 1).

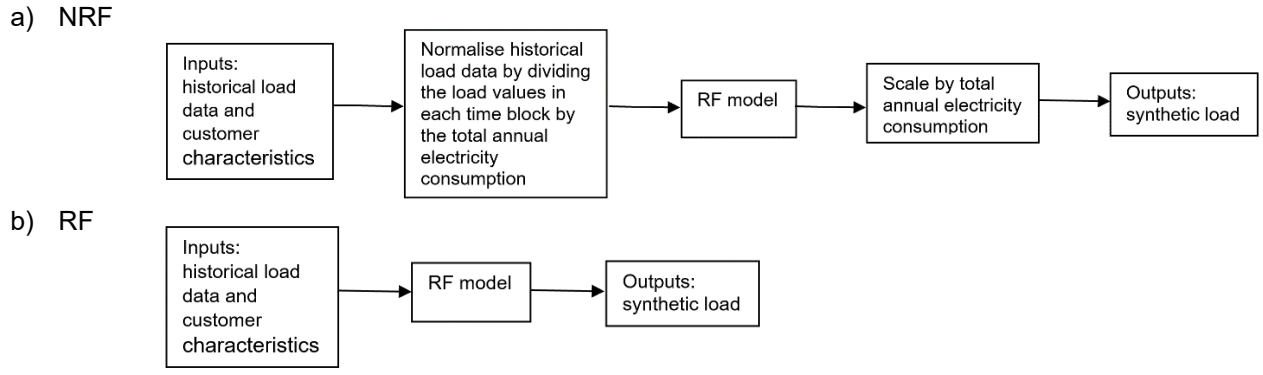


Figure 1. Methodology (a) normalised random forest (NRF) model, (b) random forest (RF) model.

We assumed daily time horizon is divided into five blocks, namely (1) 12am-7am (2) 7am-10am (3) 10am-3pm (4) 3pm-9pm, and (5) 9pm-12am. These specific time periods have been selected as they seem to represent distinct load patterns, i.e. night time with low electricity consumption, morning period, mid-day period, evening period and finally the last active period of the day. Each customer's time series values are normalised by dividing the load values in each time block by the total annual electricity consumption. Inputs to the NRF model are normalised historical load profiles (numerical input) and customer characteristics (categorical input).

Normalising the input load data is expected to enhance the RF model performance in predicting the *shape* of the load profile, rather than the size. The load profile generated could then be scaled using existing daily load benchmarks, or output from its underlying RF model to provide the predicted load profiles of the specific customers.

Feature Selection

To train the NRF model, we use the SGSC dataset, which includes over 20 survey questions about household characteristics and appliance ownership, which are the features used to train the model. Having fewer features allows machine learning algorithms to be more efficient as well as more effective, since irrelevant input features can potentially mislead machine learning algorithms and result in worse predictive performance. In feature selection, a subset of the most relevant features is therefore selected to represent a dataset.

For feature selection, we use two methods that are popular for providing unbiased results: (i) permutation-based: a feature's importance is calculated by measuring its increase in prediction error after permuting. Feature values are considered "important" if shuffling them increases model error because the model relies on them for prediction. The feature is unimportant if the model error is not affected by shuffling a feature's values, and the model ignores the feature for the prediction if the error remains unchanged. And (ii) recursive feature elimination (RFE): first, the model is fitted using all features in a given set, then each feature is removed one by one, re-fitting until we are left with the minimum necessary number of features (optimal number of features). In our selection, we select the top features that are rated as important by both methods. This enhances the reliability of the results.

Results

In this section, first, we compare the NRF with the RF model. To begin with, we compare their performance using three metrics. Next, we rank each model's features using permutation and RFE methods.

Comparison of RF and NRF models

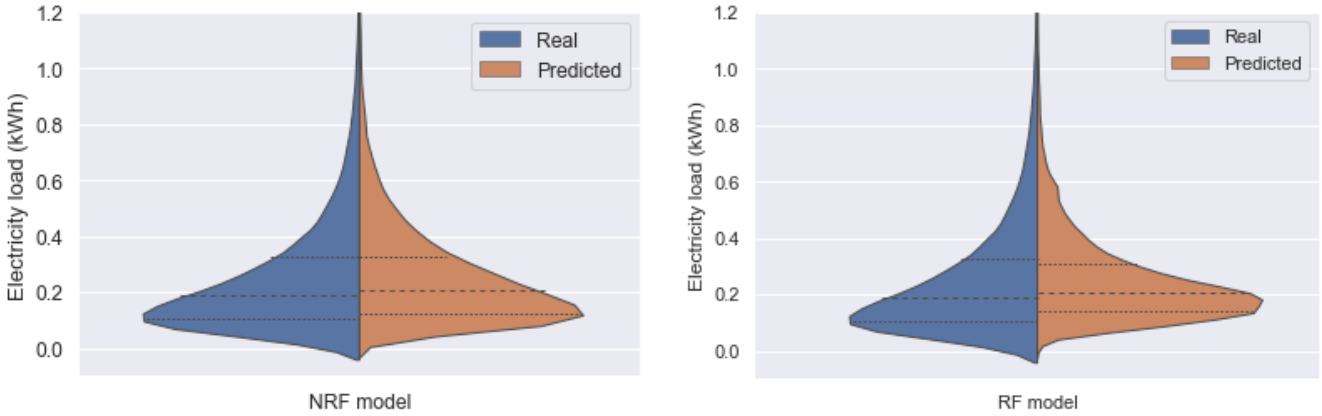
We used three metrics of (i) Pearson correlation, (ii) Spearman correlation, and (iii) mean absolute error to evaluate the NRF compared to the RF model. The results are summarised in Table 1. On both models of NRF and RF, Spearman correlation and Pearson correlation produce very close values. However, NRF shows slightly higher correlation values, indicating that real and predicted load values are sufficiently correlated. The mean absolute error between the predicted and the real load profile is normalised using an average block value from real data (shown as NMAE in Table 1). The NRF model has a lower NMAE than the RF model, so overall it outperforms the RF model.

Table 1. Comparison between RF model and NRF model

Metric	Pearson cor.		Spearman cor.		NMAE (%)	
Model	NRF	RF	NRF	RF	NRF	RF
	0.91	0.84	0.90	0.82	27.9	37.3

To compare the performance of the NRF model and the RF model, we plot the split violins of the predicted load profile versus the real load profile for each model, as shown in Figures 2a and 2b. Each model's dashed line represents its median, first quantile, and third quantile. There is a

similar discrepancy between the medians of predicted and real data for RF and NRF models. However, the first quantile and third quantile of the predicted load versus real load are very close to each other in the NRF model compared to the RF model, which supports the results that are presented in Table 1.



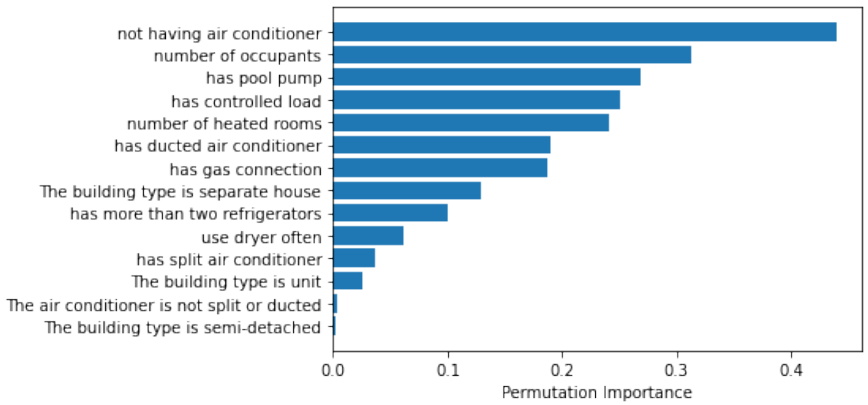
(a) NRF (b) RF

Figure 2. Comparison between the NRF model (a) and RF model (b), using violin plots.

Comparison between RF model and NRF model based on feature Importance

The SGSC data includes customer responses to over 20 survey questions. In this section, we rank the importance of each question (model's feature) using the permutation method and RFE method described above. Results are summarised in Figure 3 and Table 2 respectively.

(a) NRF



(b) RF

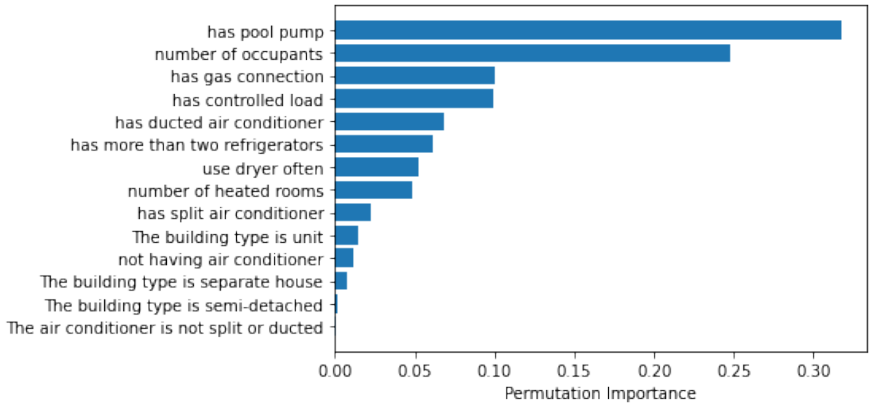


Figure 3. Feature importance using permutation method, (a) NRF, (b) RF.

It appears that applying permutation and RFE methods to the NRF model produces more meaningful results than those generated by applying the methods to the RF model. The results of the permutation method on the NRF model (Figure 3a) differ significantly from the results of the permutation method on the RF model (Figure 3b), in that the results are normally distributed when applied to the NRF model, whereas the results of using the RF model are biased towards two of the total 14 features of the models. Table 2 compares the results of the RFE method when applied to both NRF and RF models. The RFE method yields three features as the optimal number in both models. This means that by having only three features which are ranked 1 in the RFE results, the algorithms provide similar output accuracy as we feed all features into the RFE algorithm as input. As a measure, we use an average importance value of 0.15. Therefore, features with importance values greater than 0.15 are considered to be the most important features in the model. In light of this assumption, combining the results of two methods of permutation and RFE, using the NRF model, the top features are 1) absence of air conditioning, (2) number of heated rooms, (3) having a pool pump, (4) number of occupants, (5) having controlled load, (6) connected to gas, and (7) having ducted air conditioning. With the RF model, the top features include (1) having a pool pump and (2) number of occupants. It is worth pointing out that NRF model, produce results that are in line with statistics. As an evidence, large share (40%) of household energy usage in Australia is attributed to space conditioning or heating and cooling which is well reflected in the NRF model results [8].

Table 2. Feature importance using recursive feature elimination (RFE) method

Model	NRF	RF
RFE results	The optimal number of features is three.	The optimal number of features is three.
	Best features are: number of heated rooms, has pool pump, not having air conditioner.	Best features are: has pool pump, number of occupants, has controlled load.
	Rank 1: not having air conditioner	Rank 1: has pool pump
	Rank 1: number of heated rooms	Rank 1: number of occupants
	Rank 1: has pool pump	Rank 1: has controlled load
	Rank 2: The building type is separate house	Rank 2: number of heated rooms
	Rank 3: number of occupants	Rank 3: has ducted air conditioner
	Rank 4: has controlled load	Rank 4: use dryer often
	Rank 5: has more than two refrigerators	Rank 5: has gas connection
	Rank 6: has gas connection	Rank 6: has more than two refrigerators
	Rank 7: has ducted air conditioner	Rank 7: has split air conditioner

Conclusions

We proposed a random forest model with normalised input load data to generate synthetic residential load profiles based on household and appliance characteristics. Results demonstrated that by normalising the random forest model, its performance is improved as well as its usefulness for applying permutation and recursive feature elimination methods for ranking features. The model can be scaled using estimated load values from the underlying RF model or by using existing energy usage benchmarks. The synthetic load profiles generated can be used to inform household decision making around energy tariffs and deployment of rooftop solar and batteries. The improved understanding of which factors most strongly influence the shape of the household load profiles, generated through this research, can be used to inform future data collection. Our future work will involve training the model with more data, particularly datasets representing different climate conditions in different states, to improve its generality and performance.

References

[1] E. Frederiks, L. Romanach, A. Berry and L. O'Nei, CSIRO, Australia, EUDM Pilot Survey Phase 1 Postal Survey: Descriptive Results, https://near.csiro.au/assets/7a9736b1-9c78-4985-ba68-af9bd639f778, accessed: 2022-09-20 (2017).

[2] Frontier Economics, Residential energy consumption benchmarks, https://www.aer.gov.au/system/files/Residential%20energy%20consumption%20benchmarks%20-%209%20December%202020_0.pdf, accessed:2022-09-20 (2020).

[3] Smart Grid Smart City (SGSC), Customer trial data, Australian Government, https://data.gov.au/dataset/ds-dga-4e21dea3-9b87-4610-94c7-15a8a77907ef/details, accessed:2022-09-20.

[4] O. Motlagh, A. Berry, L. O'Neil, Clustering of residential electricity customers using load time series. Applied energy, 237(2019), 11–24.

[5] H. Fan, I. MacGill, A. Sproul, Statistical analysis of drivers of residential peak electricity demand. Energy and Buildings, 141(2017), 205–217.

[6] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: A review, Renewable Energy 105 (2017) 569–582.

[7] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[8] ENERGY RATING, https://www.energyrating.gov.au/products/space-heating-and-cooling#:~:text=Space%20conditioning%2C%20or%20heating%20and,and%2033%25%20in%20New%20Zealand, accessed:2022-09-20.