Applied Energy 251 (2019) 113324

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Residential battery sizing model using net meter energy data clustering

Rui Tang[a,d,*], Baran Yildiz[b], Philip H.W. Leong[a], Anthony Vassallo[c], Jonathon Dore[d]

[a] Computer Engineering Laboratory, University of Sydney, Sydney, NSW 2006, Australia
[b] School of Photovoltaics and Renewable Energy Engineering, University of New South Wales, Sydney, NSW 2052, Australia
[c] Centre for Sustainable Energy Development, University of Sydney, Sydney, NSW 2006, Australia
[d] Solar Analytics Pty Ltd, Sydney, NSW 2016, Australia

HIGHLIGHTS

- A clustering analysis on net meter energy data collected from 2779 solar households.
- K-means clustering and random forest are used for data extrapolation.
- Through clustering net meter energy data, a battery sizing model is developed.
- The proposed battery sizing model shows robustness to limited input data.
- More seasonal clusters result in more accurate battery sizing results.

ARTICLE INFO

Keywords:
Clustering
Solar energy
Smart meter
Energy storage

ABSTRACT

The high upfront costs of batteries have limited the investment in retrofit residential energy storage systems for solar customers. Battery size is one of the most important factors that impact the financial return since it determines the major operational capabilities of the solar-coupled storage system. To select the optimal battery size for a photovoltaic solar customer, it is important to perform an analysis taking account of the customer's on-site generation and consumption characteristics. However, in most cases there are insufficient pre-existing data of the required quality making it difficult to perform such analysis. In this paper, we propose a model that can achieve satisfactory battery sizing results with a limited amount of net meter electricity data. The model uses K-means clustering on customer net meter electricity data to discover important information to extrapolate limited input net/gross meter energy data and uses this in a techno-economic simulation model to determine the optimal battery size. The approach is validated using a set of 262 solar households, two tariff structures (flat and Time-of-Use) and a naive forecasting method as a comparison to the proposed model. The results indicate that the proposed model outperforms the alternative baseline model and can work with as little as one month of net meter energy data for both of the evaluated tariff structures. On average, the model results in 0.10 normalised root mean squared error in yearly battery savings and net present values, 0.07 normalised root mean squared error in annual electricity costs and a r-squared value of 0.717 in finding the optimal size of batteries. Moreover, this study reveals a linear correlation between the used clustering validity index (Davies-Bouldin Index), and errors in estimated annual battery savings which indicates that this index can be used as a metric for the developed battery sizing approach. With the ongoing rollouts of net meters, the proposed model can address the data shortage issue for both gross and net meter households and assist end users, installers and utilities with their battery sizing analysis.

1. Introduction

In recent years, driven by the technology cost reductions and government incentives, the industry has witnessed rapid rollouts of rooftop Photovoltaics (PV) systems in the residential sector. Australia leads the world in residential PV penetration, as of the end of 2015 with 15.22% of households owning a rooftop solar system [1] and this number has increased to 21.1% at the end of 2017 [2]. Several European countries also have considerable amounts of residential solar penetration, such as Belgium (7.45%), Germany (3.72%) and the UK (2.52%) [1].

Although the generous feed-in tariffs have accelerated the adoption of residential PV, most have been cancelled or reduced in various

* Corresponding author at: Computer Engineering Laboratory, University of Sydney, Sydney, NSW 2006, Australia.
E-mail address: rui.tang@sydney.edu.au (R. Tang).

https://doi.org/10.1016/j.apenergy.2019.113324
Received 24 November 2018; Received in revised form 30 March 2019; Accepted 13 May 2019

Nomenclature

η_{ch}	charging efficiency
η_d	discharging efficiency
$\beta_{j,n}$	the jth parameter for estimating $Y_{i,n}$
$\epsilon_{i,n}$	error term
b_t^{ch}	energy transferred to battery during interval t (kWh)
b_t^d	energy transferred from battery during interval t (kWh)
$bcost_t$	yearly cost after installing a battery (AUD)
c_{batt}	costs of battery, inverter in (AUD) per kWh
$c_{install}$	fixed installation costs (AUD)
C_{total}	total battery size (kWh)
$cost_0$	total capital costs including costs of battery, inverter and installation
$cost_t^{batt}$	electricity cost during interval t (AUD) for a PV system with installed battery
$cost_t^{pv}$	electricity cost during interval t (AUD) for a PV system with no battery
e_n	mean 30 min net meter energy (Wh) for an interval n
f_d	a decision tree trained by X_d and Y_d
g_t^{export}	grid export during interval t (kWh)
g_t^{import}	grid import during interval t (kWh)
$load_{energy}$	gross load energy value
m	number of intervals in one year
N	number of decision trees in the random forest model
$n_{lifetime}$	battery lifetime
net_{energy}	net energy value
$p_{season_i}^n$	cluster percentage for season i with n clusters
p_{fit}	flat feed-in tariff rate
p_{flat}	flat import tariff rate
P_{max}	rated maximum charging and discharging power (kW)
$p_{offpeak}$	off-peak import tariff rate
p_{peak}	peak import tariff rate
$p_{shoulder}$	shoulder import tariff rate
p_t^{export}	export tariff during interval t AUD/kWh
p_t^{import}	import tariff during interval t AUD/kWh
$pcost_t$	yearly cost before installing a battery (AUD)
pv_{energy}	gross PV energy value
pv_{size}	PV system size (kW)
$rate_{discount}$	discount rate
s_{code}	state code
$saving_t$	yearly saving (AUD)
$saving_{degr}$	degradation factor
SOC_{min}	minimum value for state of charge
SOC_{start}	state-of-charge when we start our simulation
soc_t	state of Charge at start of interval t (kWh)
X_d	a random subset of the training data
$X_{i,j}$	the jth feature we use for a sample i
X_{test}	vector representation of the test data
X_{train}	vector representation of the training data
Y_d	a random subset of the training labels
$Y_{i,n}$	the proportion of days clustered into seasonal cluster n for a sample i
Y_{test}	vector representation of the test labels
Y_{train}	vector representation of the training labels

countries and regions due to the reduction in technology costs [3]. Different to gross meters where all the solar generation is exported to the grid, solar generation of customers with net metering schemes is first used on site and the excess energy is then exported. Since the solar feed-in tariffs are now lower than the general import tariffs in many regions, net metering scheme is considered as a viable option to reduce the electricity costs for PV consumers. As a result, the net metering scheme has been adopted in many different countries such as Australia [4], most states in the USA [5] and Germany [6], etc. Net metering also brings opportunities to the energy storage market as batteries can now provide more benefits such as peak shaving, increasing PV self-consumption and price arbitrage.

Despite all the potential benefits that an energy storage system could offer to a net meter customer, the penetration of storage systems is still low mainly due to the high upfront costs of installing a battery system [7]. Besides, installing an energy storage system would also require a purchase of a multimode inverter to replace or add on top of the most commonly used current grid-connected inverters which further increases the upfront investments [8]. Hence, before going ahead with purchasing a battery, the financial returns or other metrics in regards to the battery capabilities need to be carefully evaluated.

Although many techno-economic simulation models have been proposed, the practicability of these approaches remains questionable due to two main reasons:

1. Many studies use synthetic household PV or load data which may result in misleading simulation results [9]. Individual households could have various consumption profiles and solar systems with different orientations, tilts or shading conditions. Moreover, it is important to use both generation and consumption data of actual solar customers as their consumption behaviours may change after the PV installations [10]
2. In order to build a robust model, a minimum amount and high quality of input PV/weather and load data is often required whereas in practice, such data might not be available.

One potential solution for the above issues is to perform data extrapolation using a customer's consumption and generation patterns extracted from the limited amount of historical data. Generally the knowledge of users' electricity consumption patterns is applied to develop tariff structure [11], demand response strategies [12], load forecasting and planning models [13,14]. Clustering is often considered as an effective tool to obtain valuable information about customer consumption behaviours and it has drawn the attention from many researchers. However, to date, the applications of this technique have mainly focused only on the electricity consumption data, ignoring the solar generation data despite the significant growth of residential solar customers. In reality, in order to gain a good understanding of the consumption and generation profiles for solar customers, it is important to conduct the clustering analysis on both the generation and consumption data.

Motivated by these facts, in this work, we introduce a battery sizing model for residential PV customers using net meter energy data clustering. The contributions of this paper are to:

1. Perform a clustering analysis on residential PV customers using net meter energy data. To the authors' knowledge, this is the first work performing a clustering analysis on net meter energy data. With the ongoing worldwide adoption of net meters, we hope our work could illustrate a new direction to load clustering research as gross load data will no longer be collected from net meter customers.
2. Present the first application of net meter clustering which is a battery sizing model that can be used for customers with net/gross meter data and is robust to limited amount of historical data and site information. To the authors' knowledge, this is the first study addressing the insufficient net meter data problem in battery size optimisation and the first work with validation of a real-time dataset and an alternative model.
3. Adopt a whole year of real-time solar and consumption data collected from a large group of 2779 solar households. To the authors' knowledge, this is the first study in the battery optimisation

literature that uses a large quantity of real-time data collected from residential solar customers. Previous studies either used much smaller or synthetic data sets for their load/solar data.

The remaining parts of the paper is structured as follows. Section 2 discusses the literature review on relevant studies. Section 3 illustrates our proposed battery sizing methodology using net meter data clustering. Section 4 presents the experimental results. Section 5 concludes this study and discusses some potential future works.

2. Literature review

Many studies perform the techno-economic analysis of PV-integrated battery systems where a lot of them focus on the battery size determinations. These studies have adopted various economic, technical and environmental indicators to be optimised in their modelling approaches [15]. The economic criteria includes levelised cost of electricity (LCOE) [16–18], net present value (NPV) [19–21], return on investment (ROI) [22] and cost-competitiveness with the grid import rate [23]. The adopted technical indicators consist of voltage deviations [18,24], energy losses [24] and frequency control [25]. Environmental criteria is generally referred to CO2 emissions where levelised CO2 equivalent life cycle emissions and damage cost of CO2 emissions are respectively considered in [18,26].

Different optimisation algorithms have been adopted to find the optimal system configuration. In [27], mixed-integer linear program (MILP) is performed to calculate the lower and upper bounds of the optimal battery sizes for a grid-connected solar system where the electricity costs stay the same when battery size exceeds the upper limit and increase significantly when the storage size is below the lower limit. MILP is also applied in a similar manner in [16,20] which optimises the system configuration and operation schedule of a PV integrated battery system. Exhaustive search is adopted in [28] to look for the battery system configuration with the lowest LCOE. A similar approach using exhaustive search is performed in [17], however the study aims for battery sizing in off-grid renewable energy systems and it also optimises battery control strategies. Stochastic mixed integer nonlinear programming (MINLP) is applied in [29] to optimise sizes and power schedules of a PV integrated battery system, with a Monte Carlo approach to model the uncertainties in PV production. A genetic algorithm (GA) [24] is applied to optimise the sizes and locations of battery-coupled distributed PV generators in distribution networks. GA is also applied in [26] to transform the optimisation cost function into a linear programming (LP) function, then the LP function is solved to find the optimal placements, sizes and power schedules in a distribution network. Reference [21] applied a dynamic programming approach to optimise sizes and energy dispatches in lithium-ion battery integrated commercial PV systems.

In summary, recent contributions in size optimisation research of PV-integrated battery system can be grouped under one of the following categories: 1. New type of optimisation criteria [23]; 2. More thorough considerations of optimisation objectives [17,18,26]; 3. Optimisation of battery control strategy or scheduling added on top of configuration determination [16,21,22,29]. 4. New battery applications in renewable energy systems [21,24]. Despite all these interesting progresses in battery optimisation models, most of the studies still use synthetic PV or consumption data. One exception, reference [19], where net meter energy data from 79 solar households is adopted, however the dataset has a considerable amount of missing data (68 days out of one year). Although synthetic solar and load profiles might still be useful when considering commercial systems, they may not be applicable for residential customers who have more diverse generation and consumption patterns. A few studies have emphasised the importance of using real-time load data in the system planning optimisation of PV battery systems. Authors in [9] compare real-time and aggregated load profiles and concludes that adopting aggregated load data may result in overestimated self-consumption and underestimated total costs. Studies like [30,31] illustrate households with various consumption patterns may result in quite different end net present values (NPVs) and self-sufficiency rate (SSR) for battery integrated solar systems. The wide adoption of synthetic data is likely due to the lack of high quality publicly available generation and consumption datasets. Moreover, even in practice, battery installers or utility who often have more direct contacts with solar customers, suffer from the insufficient data during their decision-making processes of battery system configuration. Very few research considers the limited input data problem stated above. Reference [32] uses a techno-economic model to compute the optimal PV and battery configurations for various non-solar customers and use the simulation results to develop a machine learning model that could predict the optimal configuration, NPV and SSR using a limited amount of load data. Although this model shows promising results, it still requires weather data to generate synthetic PV data. Another possible factor that could affect the practicability of the model is that a single change in the techno-economic parameters would require re-simulating and re-training of all the households in the training set.

Load clustering has been applied for many studies concerning the analysis of electricity consumption data. A couple of papers have given comprehensive reviews on the applications, techniques and evaluation

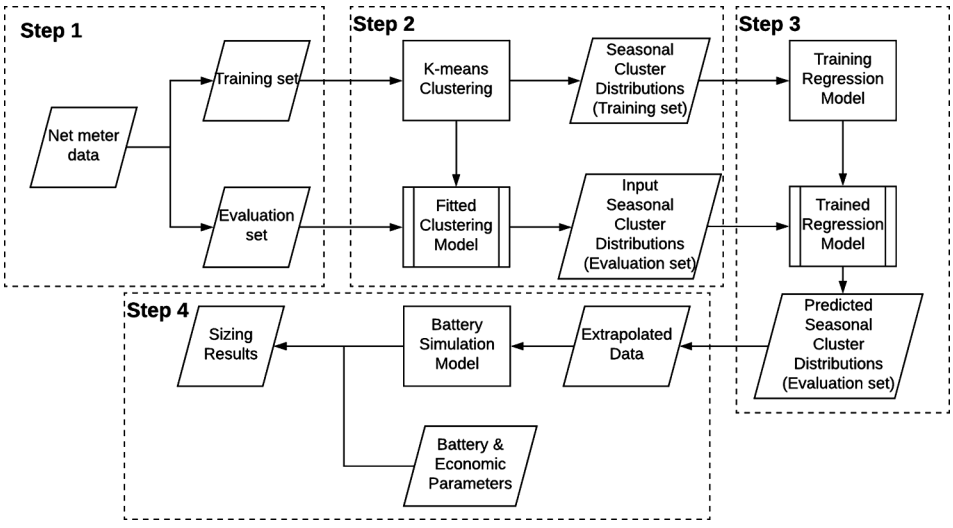

Fig. 1. The flowchart of the proposed battery sizing model using net meter clustering.

metrics of clustering [33,34]. Load profiling, which is generally referred as identification of typical consumption profiles over a certain period, is one of the main applications of load clustering and it can be used for a better understanding of consumer behaviours [35], tariff designs [11] and demand strategies [12]. Customer classification also uses load profile clustering to create cluster labels which can be related back to household characteristics and demographic information [36,37]. Moreover, load clustering has also been applied to enhance the performance of load forecasting algorithms [13,14].

A variety of load clustering techniques have been attempted in the literature such as, hierarchical clustering [13,38], k-means [35,39], fuzzy k-means [40], follow the leader [41], self-organizing map [42], support vector clustering [11] and probabilistic neural networks [43]. The number of clusters needs to be defined manually for non-hierarchical clustering models (e.g. k-means clustering) although this is not required for hierarchical and follow the leader models [34].

Various clustering validity indicators have been applied to evaluate the performances of clustering algorithms, most of them are defined using Euclidean distance metrics [33]. Commonly used clustering validity indicators (CVIs) include Clustering Dispersion Indicator (CDI) [11], Davies-Bouldin Index (DBI) [40], Mean Index Adequacy (MIA) [33], modified Dunn Index [44] and Scatter Index (SI) [11].

Clustering research taking account of consumers with on-site generation remains limited. Reference [10] applied a self-organizing map clustering model on 300 Australian households with installed PV systems which reveals a self-consumption behaviour within gross meter solar customers. A case study is demonstrated in [45] which shows how clustered consumption profiles can be used for the size planning of a PV and energy storage system on a commercial building.

3. Methodology

The main purpose of this analysis is to develop a model for solar customers with net meter arrangements and limited amounts of historical consumption and generation data in order to decide on the most optimal battery size. The methodology (shown in Fig. 1) can be separated into four parts:

1. We prepare our dataset for net meter clustering and separate it into a training set and an evaluation set. The training set is used for fitting the parameters of the clustering and regression models mentioned below and the evaluation set is used to evaluate the performance of the proposed model.
2. K-means clustering is applied to cluster net meter energy data separately for various seasons; summer, autumn, winter and spring. For each household, we determine the seasonal cluster distributions which specify each household's percentages of seasonal net meter profiles partitioned into each seasonal cluster.
3. Regression models are trained on the obtained seasonal cluster distributions and used to extrapolate the seasonal cluster distributions of the new input net meter energy data at a given length.
4. The extrapolated data, battery and economic parameters are fed into a battery simulation model which produces the optimal battery sizes for the customers in the evaluation set.

For comparing our approach, an alternative naive prediction method (see Section 3.4.1) is also implemented. To evaluate the performances of the two methods, the battery sizing results are also derived for the ideal case where a whole year's real data is provided to the battery simulation model instead of extrapolated data. This allows us to determine errors in the battery sizing results for the two implemented modelling approaches.

3.1. Step 1 – pre-clustering step

3.1.1. Data collection

The data used in this research was collected by Solar Analytics, an Australian solar monitoring company [46] using Wattwatchers monitoring hardware [47] that monitors both solar generation and electricity consumption. The dataset includes 5-min gross PV and consumption data collected between December 2016 and December 2017 from 2779 Australian solar households. By using solar and load data collected from the same households, we take account of potential impacts of domestic solar generation on consumption behaviours. We ensure these customers have adequate amounts of data: the overall amount of missing data is less than 3% and the customer with the most missing data has 7% of data missing. The DC solar system ratings of these customers are also recorded and we ensure these rooftop PV systems have been performing normally without any major system faults within this period. For this study, in order to construct a net meter dataset from gross meter data, we first convert the gross PV and consumption data to net meter energy data using (1) and then resample the net meter energy data to 30-min temporal resolution.

$$net_{energy} = pv_{energy} - load_{energy} \quad (1)$$

Before applying any clustering, load curve normalisation has been applied in some previous load clustering studies [37,43,48,49], on the other hand some consumption clustering studies just use raw consumption data [10,42,50]. In this study, after carrying various simulations, using normalised data did not produce as good results as the raw data so it was decided to present results for only the raw data.

3.1.2. Data split

In order to properly evaluate our battery sizing model, we split our dataset into a training set and an evaluation set. Clustering is only performed on the training set which includes 2517 randomly-selected customers, the remaining 262 households included in the evaluation set were treated as new customers to evaluate the robustness of the proposed battery sizing model against limited input data.

3.2. Step 2 – net meter clustering

As seasonality generally exists in both solar generation and consumption data, we divide our dataset into four seasons and perform clustering on each of them. Four seasons are defined as follows for Australia [51]: Summer: December to February, Autumn: March to May, Winter: June to August, Spring: September to November. Each daily profile of the customers in the training set was used in clustering so that most information is captured during the clustering process.

3.2.1. Clustering algorithm

We use K-means algorithm [52] for the net meter profiles as it has been proven to be simple yet effective in previous load clustering studies [37,49] and furthermore it converges quickly which is a great advantage for large datasets [53].

3.2.2. Clustering evaluation

In the previous clustering studies [11,42], clustering validity indicators (CVIs) have been used to evaluate the performance of consumption data segmentations and to find the optimal number of clusters. On the other hand, the end use application of this study is choosing the optimal battery size and approximating potential savings, therefore, the number of clusters were chosen according to the minimum errors obtained for these tasks. In the mean time, in order to see whether there is any relationship between the CVI and errors obtained in battery sizing results, Davies-Bouldin index (DBI) [54] is also computed.

3.2.3. Seasonal cluster distributions

After separating the training data into four seasons, daily net meter profiles are clustered into various seasonal clusters. As a result, the distributions of each household's clustered net meter profile are calculated for these seasonal clusters. They simply describe each household's percentages of seasonal net meter profiles assigned to each seasonal cluster. It should be noted a seasonal cluster distribution does not require the whole season of data to be computed, in fact, it can be calculated for any period within the season. For instance, when a new customer has 30 days of net meter energy data in summer where 18 daily profiles are grouped into cluster 1 and 12 days are in cluster 3. Then the summer cluster distribution of this household is 60% (18/30) in cluster 1, 40% (12/30) in cluster 3 and 0% for other seasonal clusters.

Seasonal cluster distributions reveal the typical seasonal net meter patterns and their occurrences for a household at a given period. Therefore, when two households have similar seasonal cluster distributions within a period, they show similar net meter profiles in the same period. Moreover, the seasonal cluster distributions can be used as extracted features to predict seasonal distributions of other unknown periods which will be shown in following sections.

3.3. Step 3 – predict seasonal cluster distributions

This step trains a machine learning model to predict the seasonal cluster distributions for new customers with limited net meter energy data. In this work, multivariate linear regression and random forest regression techniques were compared. Feature selection and hyper-parameter tuning are adopted to enhance the performance of regression models. The main steps to train the machine learning model are shown in Fig. 2. For both regression techniques, feature selection is applied using the regression model with default hyper-parameters and then parameter tuning is performed to select hyper-parameters that lead to superior regression results, after that the tuned model and selected features are used for model training. Finally, after training, the trained model is used to predict seasonal cluster distributions. For each predicted season/period, we search for a customer which shows the most similar seasonal cluster distributions and has full length of data. In particular, this is done by finding a customer in the training set with the shortest Euclidean distance in terms of seasonal cluster distributions in the predicted season/period. This customer's data is then used as the seasonal extrapolated data for the new customer.

In terms of the length of data from the new customers, we experiment three options; a single month, a single season or two random seasons. Also to evaluate the impacts of applying different months/seasons as inputs, we test all the input data scenarios in Table 1. Finally, the model output values are the seasonal cluster distributions for the remaining period of a year.

3.3.1. Features

The input features used for predicting seasonal cluster distributions are listed in Table 2, which contains each household's: DC solar system size, state code, the daily averaged 30 min net meter energy and the seasonal cluster distributions of the net meter energy data for the given period. To compute the averaged daily net meter energy, we simply calculate the averaged energy within the known data period for each 30 min interval of a day. Our dataset includes customers from 6 Australian states/territories: Australian Capital Territory, New South Wales, Queensland, South Australia, Victoria and Western Australia. They are converted to integers from 1 to 6.

If the input data has overlapping period between different seasons, for example if the given input data has 60 days which include 20 winter days and 40 spring days, the input seasonal cluster percentages would include both winter and spring cluster percentages for the provided data. The predicted values would be the seasonal cluster distributions in summer, autumn and the remaining periods of winter and spring.

Table 1
Tested input data scenarios.

Input Data Length	Tested Input Data Scenarios
One month	Month number in {1–12}
One season	Seasons in [Summer, Autumn, Winter, Spring] {1–4}
Two seasons	Two-season combinations in {1&2, 1&3, 1&4, 2&3, 2&4, 3&4}

Table 2
Features used for regression.

Feature Name	Symbol
seasonal cluster percentages of season i with n clusters	$p^1_{season_i}$, $p^2_{season_i}$, ...,$p^n_{season_i}$
mean 30 min net meter energy (Wh)	e_1, e_2, ...,e_{48}
PV system size (kW)	pv_{size}
state code	s_{code}

3.3.2. Multivariate linear regression

Multivariate linear regression [55] is a machine learning approach where multiple independent variables are used to predict multiple dependent variables. The regression problem in this study can be formulated by (2) using this technique. The ordinary least square method which minimises the squared differences between training labels and predicted values, is applied to estimate the parameters in (2).

$$Y_{i,n} = \beta_{0,n} + \beta_{1,n}X_{i,1} + \beta_{2,n}X_{i,2} + \beta_{3,n}X_{i,3} + \cdots + \beta_{p,n}X_{i,p} + \epsilon_{i,n} \quad (2)$$

where $Y_{i,n}$ is the predicted proportion of days clustered into seasonal cluster n for a sample i, $X_{i,j}$ is the jth feature we use for a sample i, $\beta_{j,n}$ is the jth parameter for estimating $Y_{i,n}$ and $\epsilon_{i,n}$ is the error term. In this paper we use the Python implementation of this model [56].

3.3.3. Random forest regression

Random Forest (RF) is an ensemble machine learning method which trains multiple decision trees on different random subsets of the training data [57]. The adopted RF model uses bootstrap aggregating (bagging) technique for training, where random subsets are drawn with replacements and each subset has the same sample size as the original training set [58]. Given the training data X_{train} and the output label Y_{train}, by using bagging, N random subsets are generated from X_{train} and

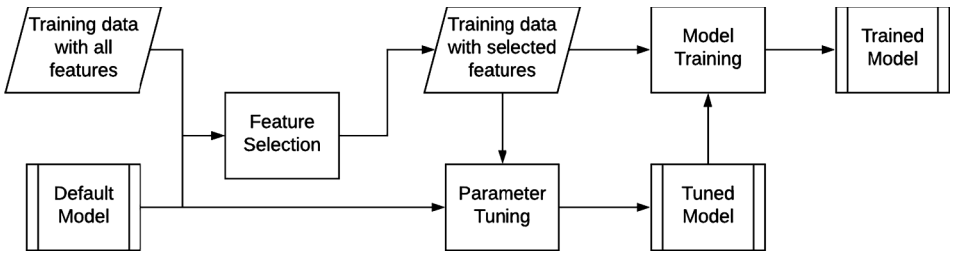

Fig. 2. Main steps of the regression model training.

Y_{train} (denoted as X_d and Y_d). For each sampled subset, we train a decision tree f_d using X_d and Y_d. Then when making a prediction for a new sample after training, the RF model will aggregate the predictions from these decision trees. For regression tasks, the aggregation function takes the mean of the predictions by various decision trees (shown below in (3)).

$$Y_{test} = \frac{1}{N}\sum_{d=1}^{N} f_d(X_{test}) \tag{3}$$

where Y_{test} denotes predicted labels of the test set, X_{test} is the input test data.

As a result, compared to a single decision tree trained with the whole dataset, RF generally has a better performance as it reduces the model variance whilst resulting in similar bias errors [59]. In this study we use the Python implementation of this model [56] and we set N to 100 which means results from 100 decision trees are aggregated within the RF model.

3.3.4. Feature selection

For the linear regression model, we applied a Lasso (Least Absolute Shrinkage and Selection Operator) regression analysis [60] which performs both L1 regularisation and feature selection. It penalises the absolute sum of coefficients, as a result, the regression coefficients for some features shrink towards zero and hence they are filtered out from the model.

The Boruta algorithm [61] is applied to select features for the RF model as it previously outperformed other RF feature selection approaches [62]. The algorithm randomly performs permutation on all features and train the RF model using both the original and shuffled features [61]. For each original feature, a statistical test is conducted which computes the confidence towards a better importance value compared to the maximum importance value of the shuffled features. Features with significantly higher importances are marked as important features whereas features with smaller importances are removed. Then the process will re-iterate until all features are categorised as confirmed/rejected or until a certain number of iterations is reached. In this study, we used the Python implementation of Boruta [63] and set the maximum number of iterations to 30. It is also suggested in [63] that the original threshold where a real feature needs to have better importance than all the shuffled features can sometimes be too stringent so we set the percentile parameter to 80% which means true features will pass the statistical test when its importance is higher than 80% of the shuffled features.

3.3.5. Parameter tuning

In order to achieve better performances from our regression models, we optimise the hyper-parameters of the models by using random search along with 10-fold cross validation (CV). Compared to other hyper-parameter optimisation approaches such as grid search and manual search, random search has proven to be more efficient in terms of computational costs [64]. The hyper-parameters tuned for the linear regression and RF models are shown in Table 3. Some of the default parameters are selected from their default values set by sklearn [56] and the others are selected by experience to create a loosely tuned default model for feature selection.

In a 10-fold CV, it randomly splits the training set into 10 equal sized subsets. 9 subsets are used as training data and the remaining subset is evaluated once as a test set. This validation process is repeated 10 times where each time a different subset is used as a test set, after that we compute the averages and standard deviations of the mean squared error (MSE) in seasonal cluster proportions. We then choose the hyper-parameters that yield the lowest averaged MSEs.

3.4. Step 4 – battery sizing model

After predicting the seasonal cluster distributions for all the listed input data scenarios in Table 1 and extrapolating the net meter energy data for the unknown period, we determine the battery sizing results by feeding the extrapolated net meter profiles for the entire year to a battery simulation model which is described below in detail.

3.4.1. Alternative approach for data extrapolation

For comparing our net meter clustering approach, an alternative method is adopted. In this method, instead of performing clustering on the new customers, a naive forecasting approach is applied which finds another customer from the training data that has the most similar net meter profile in the known period, measured by finding the shortest Euclidean distance between net meter profiles. Then for predicting remaining periods of the year for the new customer, we just use the net meter energy data of the closest site as a naive prediction. Furthermore, in order to evaluate the performances of these two prediction models, the battery sizing results are also derived for the ideal case; where a whole year of real monitored net meter energy data is provided to the battery simulation model. This ideal case allows us to compute the errors in optimal battery sizes, net present values, yearly battery savings and electricity costs for the net meter clustering case and the naive forecasting approach.

To properly assess these three approaches, the battery sizing results are only computed for the evaluation set as it has not been used for fitting the parameters of the clustering and regression models.

3.4.2. Model parameters

Key parameters used in the battery simulation model are listed in Table 4. The model simulates annual realistic battery operations and computes battery sizing results using the listed battery and economic

Table 3
Tuned Parameters for Lasso model and RF model [56].

Parameter in sklearn	Parameter Description	Tuning Range	Default Value
Lasso Regression Model			
alpha	constant to multiply the L1 regularisation term	float in {0–1}	1.0
RF Model			
max_features	the number of randomly drawn input features when considering the best split	n, $\sqrt{n}$, or 1/3n (n is the number of all features)	n
max_depth	The maximum depth of the decision trees	integer in {2–12}	6
min_samples_leaf	The minimum number of required samples at a leaf node	integer in {1–12}	2
min_samples_split	The minimum number of required samples to make an internal split	integer in {2–12}	2

Table 4
Battery simulation parameters.

Parameter	Definition	Values
Battery Specifications		
C_{total}	Total Battery Size (kWh)	1–15 kWh
P_{max}	Rated maximum charging/discharging power (kW)	0.4 × C_{total}
SOC_{min}	Minimum value for state of charge	20%
SOC_{start}	SOC when simulation starts	0%
η_{ch}	Charging efficiency	90%
η_d	Discharging efficiency	90%
Economic Parameters		
$n_{lifetime}$	battery lifetime	15 years
$rate_{discount}$	discount rate	0.03
$saving_{degr}$	yearly reduction in saving due to battery degradation	0.05
Tariffs (in AUD/kWh)		
p_{flat}	flat import tariff rate	$ 0.30/kWh
p_{peak}	peak import tariff rate	$ 0.45/kWh
$p_{shoulder}$	shoulder import tariff rate	$ 0.25/kWh
$p_{offpeak}$	off-peak import tariff rate	$ 0.15/kWh
p_{flt}	flat feed-in tariff rate	$ 0.11/kWh

parameters for the three approaches discussed above: our net meter clustering approach, the naive prediction method and the ideal case where the whole year's data is provided.

3.4.3. Rule-based (RB) model

We assume the battery charging/discharging follows a simple rule-based algorithm in Algorithm 1 that has the main objective of maximizing solar self-consumption. This modelling approach has previously been implemented for some studies [65,66] and also used in practice at many installed battery sites due to its simplicity and ease of implementation.

Algorithm 1. Pseudo Code for the rule-based model

Input net_energy_t, P_{max}, C_{total}, SOC_{min}, p_t^{import}, p_t^{export}, SOC_{start}
▷Input parameters
Input η_{ch}, η_d ▷ Import charging/discharging efficiencies
$soc_o \leftarrow SOC_{start}$ ▷Set initial SOC to SOC_{start}
for t in (1, 2, …, m)**do** ▷ for loop starts
  $soc_t^{usable} \leftarrow C_{total} \times (1 - SOC_{min})$
  if $net\_energy_t > 0$ **then** ▷when there is excess solar, charge battery until full
    $b_t^{ch} \leftarrow min(net\_energy_t, P_{max}, (soc_t^{usable} - soc_{t-1})/\eta_{ch})$
    $g_t^{export} \leftarrow net\_energy_t - b_t^{ch}$
    $cost_t^{pv} \leftarrow -net\_energy_t \times p_t^{export}$
    $cost_t^{batt} \leftarrow -g_t^{export} \times p_t^{export}$
    $soc_t \leftarrow soc_{t-1} + b_t^{ch} \times \eta_{ch}$
  else ▷when there is excess demand, discharge battery until depleted
    $b_t^{d} \leftarrow min(-net\_energy_t, P_{max} \times \eta_d, soc_{t-1} \times \eta_d)$
    $g_t^{import} \leftarrow -net\_energy_t - b_t^{ch}$
    $cost_t^{pv} \leftarrow -net\_energy_t \times p_t^{import}$
    $cost_t^{batt} \leftarrow g_t^{import} \times p_t^{import}$
    $soc_t \leftarrow soc_{t-1} - b_t^{d}/\eta_d$
Output $\sum_{t=1}^{m} cost_t^{pv}$ $\sum_{t=1}^{m} cost_t^{batt}$

3.4.4. Determine optimal battery size

We determine the optimal battery size by searching for the value which maximises the Net Present Value (NPV) at the end of the battery lifetime, shown below in (4). The current residential batteries in the market is between 1 and 15 kWh [67] so we use this range for our grid search (i.e. 16 values in total including 0 kWh which means no battery is installed). The averaged warranty provided by manufacturers is around 10 years [67] however adopting a 10-year lifetime makes it infeasible to install batteries for the majority of solar customers even with a low battery price scheme. We therefore adopt 15 years for the maximum lifetime of a battery in our simulation model so it would be easier to compare errors in optimal battery sizes for the two tested approaches.

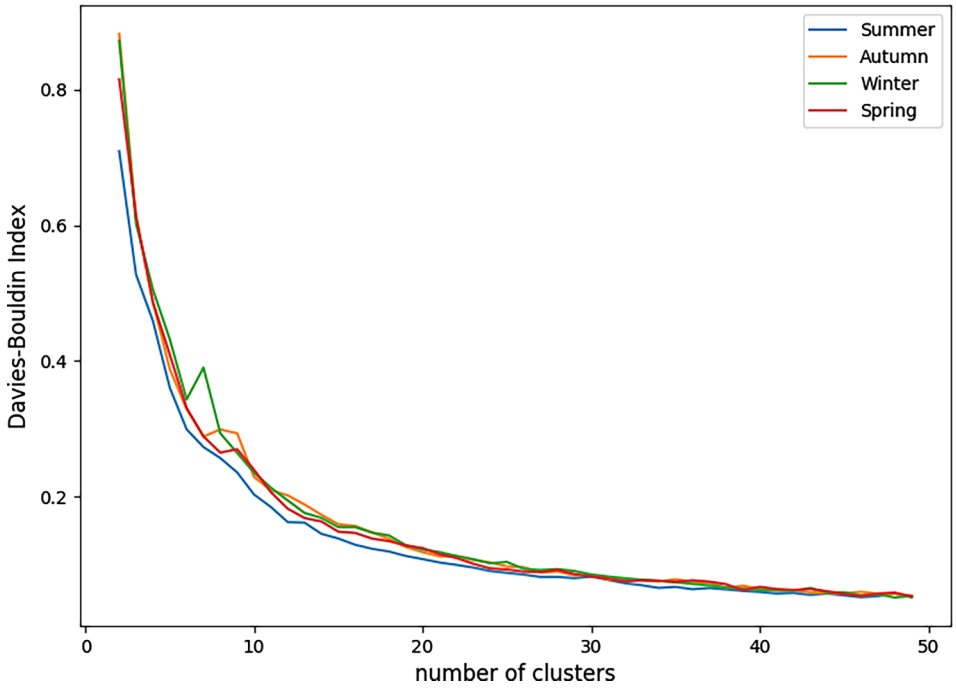

Fig. 3. Davies-Bouldin Index for adopting various numbers of clusters each season using raw data.

$$NPV = -cost_0 + \sum_{t=1}^{n_{lifetime}} \frac{saving_t \times (1 - saving_{degr})^t}{(1 + rate_{discount})^t}$$
$$= -(c_{batt} \times size_{batt} + c_{install}) + \sum_{t=1}^{n_{lifetime}} \frac{(pcost_t - bcost_t) \times (1 - saving_{degr})^t}{(1 + rate_{discount})^t} \quad (4)$$

where $cost_0$ is the total capital costs including costs of battery, inverter and installation. We assume costs of a battery and a new multimode inverter increase by c_{batt} when adding 1 kWh of battery capacity and installation costs ($c_{install}$) remain the same. $saving_{degr}$ is a degrading factor on yearly battery savings, we assume savings will reduce annually by 5% due to battery degradation to save our computational costs, this is an arbitrary estimated parameter determined by the general guaranteed end lifetime usable capacity which is 60% $\approx (1 - 5\%)^{10}$ [68]. Yearly saving ($saving_t$) is simply derived by subtracting the yearly cost ($pcost_t$) without installing a battery and annual costs after the battery installation ($bcost_t$).

4. Results and discussion

4.1. Clustering results

For each season, the Davies-Bouldin Index (DBI) is calculated for adopting various numbers of clusters to cluster the training set of 2517 customers where a smaller value of DBI indicates better clustering outcome. As shown in Fig. 3, seasonal DBIs improve as the numbers of seasonal clusters increase. However, as making too many clusters could result in clustering results that are not desirable for the post clustering applications hence user inspection is often required. Authors in [34] suggested locating the "elbow points" in a DBI curve as the numbers of clusters in terms of segmentation quality since DBI improves little beyond these points. We adopted the same approach in this work, as a result, Fig. 4 illustrates the seasonal cluster centroids using the optimal numbers of seasonal clusters determined by DBI.

For Summer clustered groups, cluster 1 and 5 have similar peaks of grid import and export whereas in cluster 4, evening load is much higher than the export around noon. Customers who have majority of the net meter profiles in cluster 3, 8 and 10 have higher solar generation compared to night-time and early morning consumption. Electricity import and export are both at low levels in cluster 2, on the other hand in cluster 6, 7 and 9, on average there is no export mainly due to higher levels of daytime consumption. Overall, net meter profiles in cluster 2, 3, 8, 10 are more likely to benefit from small-size batteries as their have low level imports whereas larger batteries are more suitable for profiles in cluster 1, 4, 5 which have considerable amounts of imports and exports. For cluster 6, 7 and 9, energy storage is not a good option as on average there is no excess PV generation.

In Autumn, high export and low night time grid import is observed in cluster 4, 5 and 9 where cluster 4 has a lower export compared to cluster 5 and 9. Cluster 2, 3, and 7 all show considerable amount of import and export however cluster 3 has a higher night time load compared to the other two. Centroids of cluster 6 and cluster 8 both have zero net export with a morning and a evening consumption peaks whereas cluster 1 has small amounts of import and export. Small-size batteries seem to be beneficial for net meter profiles in cluster 1, 4, 5, 9 where small amount of energy is required from the battery to cover the consumption in non-solar periods. cluster 2, 3, and 7 will get more savings from larger battery sizes whereas it would be hard to utilise batteries for profiles in cluster 6 and 8 as there is no excess generated energy.

For Winter, a few groups (cluster 4, 8 and 9) have low night-time consumption and noticeable amounts of exports, on the other hand three cluster centroids (5, 7, 11) have zero net generation. Low export and high import is observed in cluster 1, 3, 6 and 10 where cluster 1 and 3 show higher night-time consumption while cluster 6 and 10 have higher morning load. Relatively high export and import are shown in cluster 2. Overall, most net meter profiles can only utilise a small amount of battery capacity as they either have low net consumption (cluster 4, 8 and 9) or their generation is low (cluster 1, 3, 6 and 10). Net meter profiles in cluster 5, 7 and 11 have insufficient energy to charge batteries whereas cluster 2 can fully utilise a medium or large size residential battery.

In Spring, centroids of cluster 1, 4 and 10 show zero export, however cluster 2, 3 and 9 have significant grid exports. Three clusters (5, 6 and 8) have considerable exports and imports whereas the levels of import and export are both low in cluster 7. Small batteries can be more beneficial for net profiles in cluster 2, 3, 7 and 9 whereas larger batteries can be fully utilised for cluster 5, 6 and 8. On the other hand, energy storage cannot be utilised at all in cluster 1, 4 and 10.

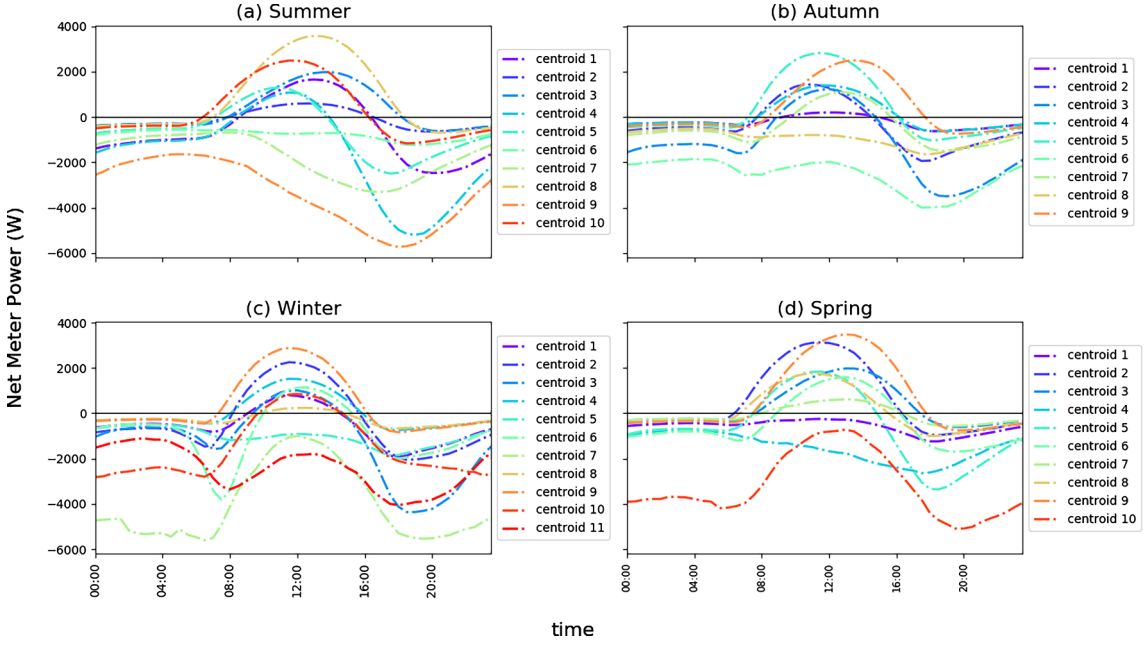

Fig. 4. Seasonal unnormalised cluster centroids in (a) Summer, (b) Autumn, (c) Winter, and (d) Spring using optimal numbers of clusters.

4.2. Regression results

4.2.1. LR model vs RF model

Figs. 5–7 illustrate the mean squared errors (MSEs) in predicted seasonal cluster proportions for various evaluated input data lengths using 5 seasonal clusters for each season as an example. 10-fold cross validation is performed to generate MSEs for each randomly selected subset of the training set, this allows us to generate boxplots to display the distributions of MSEs. The results indicate Random Forest (RF) model outperforms the Multivariate Linear Regression (MLR) model for all the evaluated scenarios therefore we will adopt this model for the data extrapolation process.

When using monthly or seasonal data as input, Autumn tends to produce the best regression results and has much smaller MSEs compared to the scenarios using Winter or Summer. This is likely due to the fact that Autumn has a more balanced generation and consumption whereas winter and summer have either dominant generation or consumption.

For predicting new households net meter profiles by applying single monthly data as inputs; Winter seasonal cluster proportions seem to be the hardest seasonal cluster distributions to predict while it is much easier to determine these values in Autumn and Spring. This is likely caused by low irradiance in winter which causes the winter cluster distributions to be heavily influenced by household consumptions whereas the solar generation is more dominant within the input data period.

January seems to be the worst month for predicting other seasons as it generates the highest MSEs in predicted cluster proportions. It is interesting to note that to predict cluster distributions in Spring, April produces the best results whereas for other three seasons, the months adjacent to the predicted seasons have the lowest MSEs.

It is also interesting to note that in some cases when months adjacent to the predicted seasons are used (e.g. using May to predict cluster distributions in Winter), predicting with one month of data results in lower MSEs compared to one whole season of input data. The reason for that is probably months adjacent to the predicted season have quite similar consumption and generation patterns to the predicted season, adding other months actually results in worse input features (i.e. the seasonal cluster distributions and mean net meter energy values).

4.2.2. MSE vs number of clusters

For each input data length, we average the MSEs for each tested scenario and plot them against various number of seasonal clusters. The same number of clusters are applied in each season to avoid creating too many combinations. As shown in Fig. 8, the RF model still outperforms the LR model when using other numbers of seasonal clusters. MSEs in predicted seasonal cluster distributions are reduced when we increase the number of clusters in each season. After the number of seasonal clusters reaches 30, the improvements in the averaged MSE slow down significantly.

4.2.3. Feature selection and parameter tuning

Feature selection and parameter tuning both have improved the regression results. For example for the specific case where we input data in summer to predict seasonal cluster distributions in Spring and 5 clusters are used for each season. 42 features are selected after applying the Boruta algorithm on the default RF model, then parameter tuning is performed. As a result, compared to the original RF model with default features that produced 10-fold cross-validation MSE of 0.01412 (mean) ± 0.00154 (standard deviation), the MSE derived after feature selection and parameter tuning is 0.01389 (mean) ± 0.00139 (standard deviation).

4.3. Battery sizing results

4.3.1. Errors in yearly costs and savings

Figs. 9 and 10 show the normalised root mean square errors (NRMSEs) in yearly savings and costs against the number of seasonal clusters for both naive forecasting case and net meter clustering case with different input data lengths. The evaluated range for the number of seasonal clusters is from 3 to 40, we use equivalent numbers of clusters for each season to avoid creating too many combinations for our analysis. Various battery sizes range from 1 to 15 kWh along with different input data scenarios listed in Table 1 are all tested and averaged for each analysed number of cluster. For the plot labels, we use prefix “net” and “naive” to represent the two tested methods: the net meter clustering approach and the naive forecasting method. The suffix is used to differentiate various input data lengths (i.e. “one_season” indicates applying one season of input data to extrapolate data in other seasons). It should also be noted that Fig. 10 illustrates errors for two types of costs, one is the yearly electricity costs before installing a battery (“pre_cost”) and the other one is the costs after installing a battery (“batt_cost”).

By comparing errors in yearly savings, it is clear that the net meter clustering approach outperforms the naive method for both Time-of-Use (ToU) and flat tariffs. Especially for the flat tariff case, using one month of data and the net meter clustering model actually produce

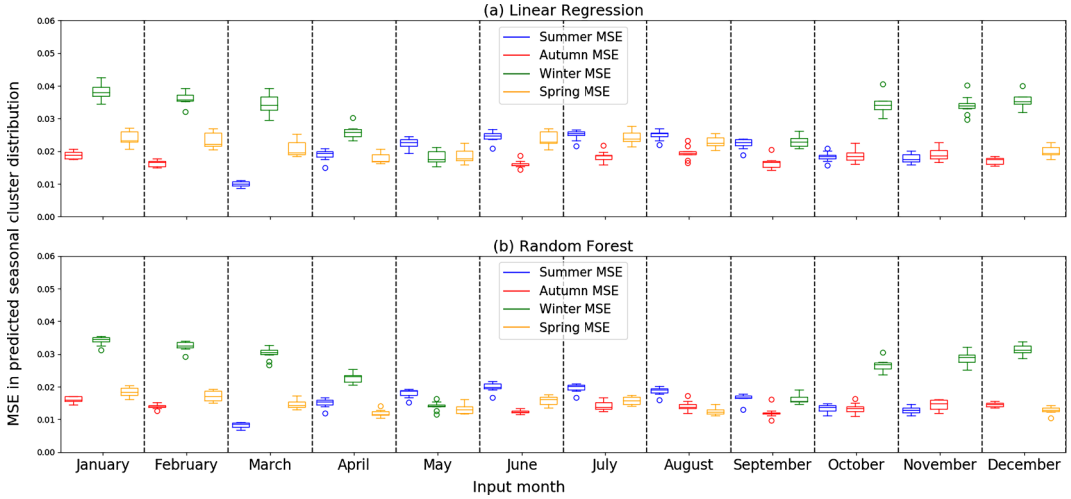

Fig. 5. Mean Squared Error (MSE) in predicted seasonal cluster distributions using one month of input net meter energy data for the adopted (a) linear regression model, (b) random forest model.

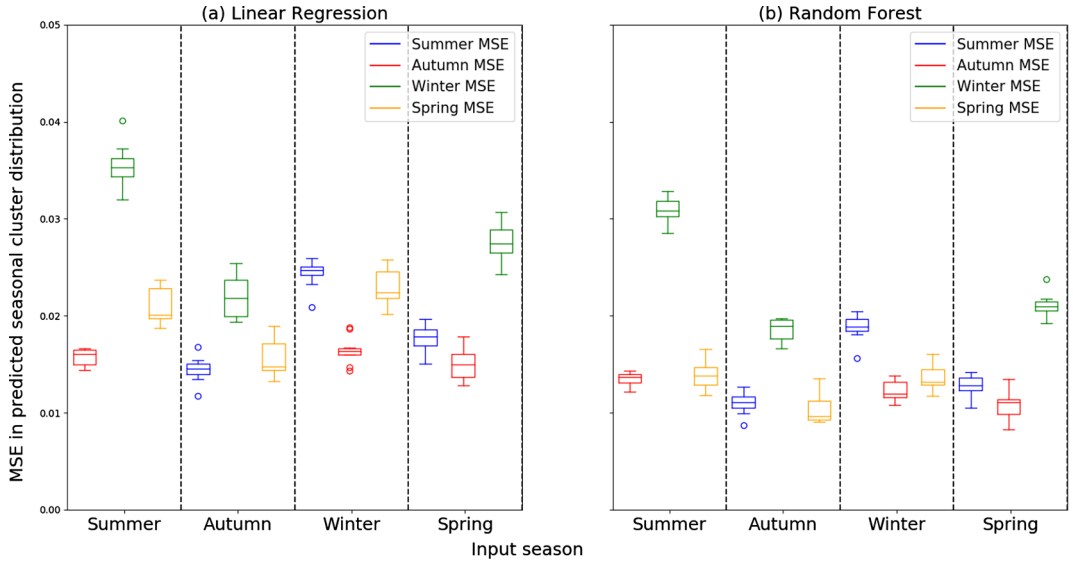

Fig. 6. Mean Squared Error (MSE) in predicted seasonal cluster distributions using one season of input net meter energy data for the adopted (a) linear regression model, (b) random forest model.

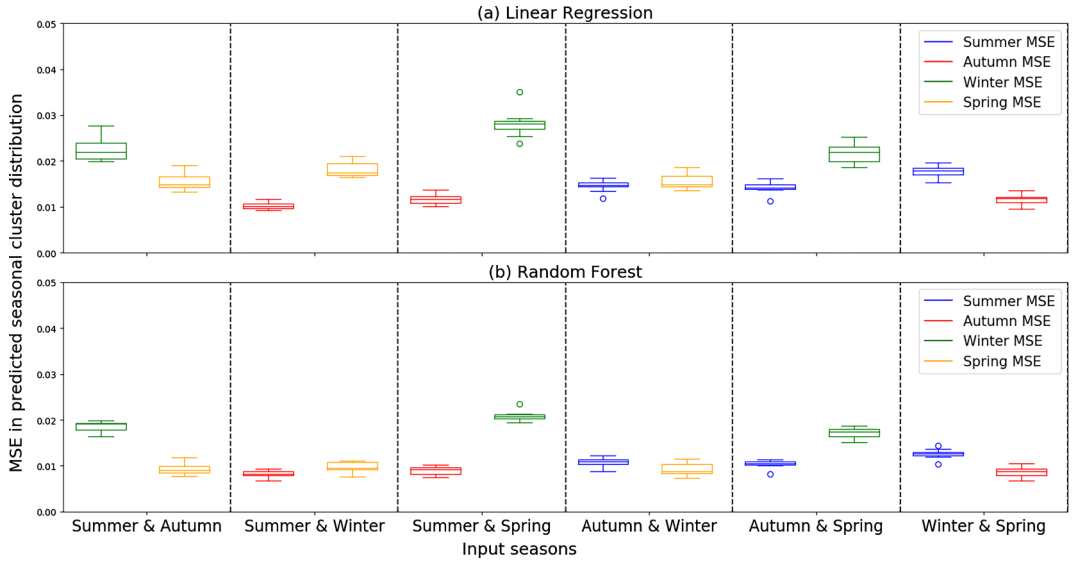

Fig. 7. Mean Squared Error (MSE) in predicted seasonal cluster distributions using two seasons of input net meter energy data for the adopted (a) linear regression model, (b) random forest model.

smaller errors than applying one season of data with the naive forecasting approach for all the evaluated numbers of seasonal clusters. Another obvious trend is that as the number of seasonal clusters increase, the NRMSEs in yearly savings are reduced for all the analysed input data lengths. Moreover when low numbers of clusters are adopted for net meter clustering method, the errors in savings are lower for flat tariff compared to ToU however as the number of clusters increases, the NRMSE drops more quickly for ToU. As a result, at high number of seasonal clusters, they both have similar NRMSEs in estimated yearly savings.

Errors in yearly costs seem to present similar trends as the errors in savings, the net meter clustering approach tends to have much smaller NRMSEs in yearly electricity costs before and after installing batteries and the differences between the net meter clustering approach and the naive method get larger when we increase the number of seasonal clusters. When the net meter clustering model is applied, one month input data outperforms the naive forecasting method using one season of data for both tested tariff structures and applying one season input data result in similar NRMSEs as the naive forecasting approach with two seasons of input net meter energy data.

This means by applying net meter clustering, we could make better estimations in yearly electricity costs and battery savings when a limited amount of net/gross meter data is provided. Not only this can improve the battery sizing procedures of installers or utility, potentially it can also better assist the end-users to select the best tariff offers to reduce their energy costs with a small amount of historical data for their home energy systems. As shown in Fig. 10, the NRMSEs in costs before and after installing a battery are both much lower using the net meter clustering approach for both evaluated tariff structures. Therefore, the solar customers could apply different tariff structures on their data extrapolated by the net meter clustering model and expect much smaller errors in estimated electricity costs compared to the baseline naive forecasting method, regardless of whether future battery purchase decisions are considered.

Another aim of the study was to explore whether the DBI is correlated to the battery sizing results. Fig. 11 shows the errors in yearly battery saving against averaged seasonal DBI values. We can see a linear correlation between DBI and NRMSEs in yearly savings for all the

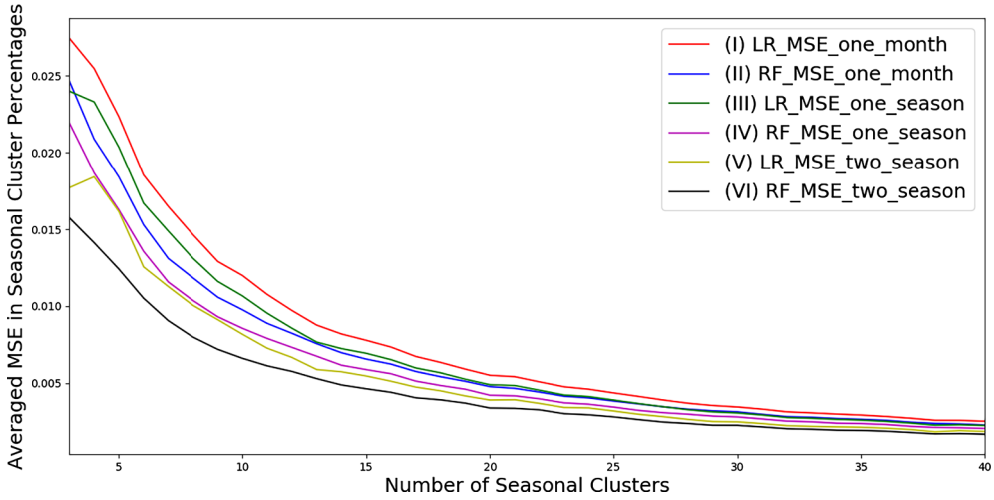

Fig. 8. MSEs vs no seasonal clusters when applying (I) one month of input data and LR, (II) one month of input data and RF, (III) one season of input data and LR, (IV) one season of input data and RF, (V) two seasons of input data and LR, and (VI) two seasons of input data and RF.

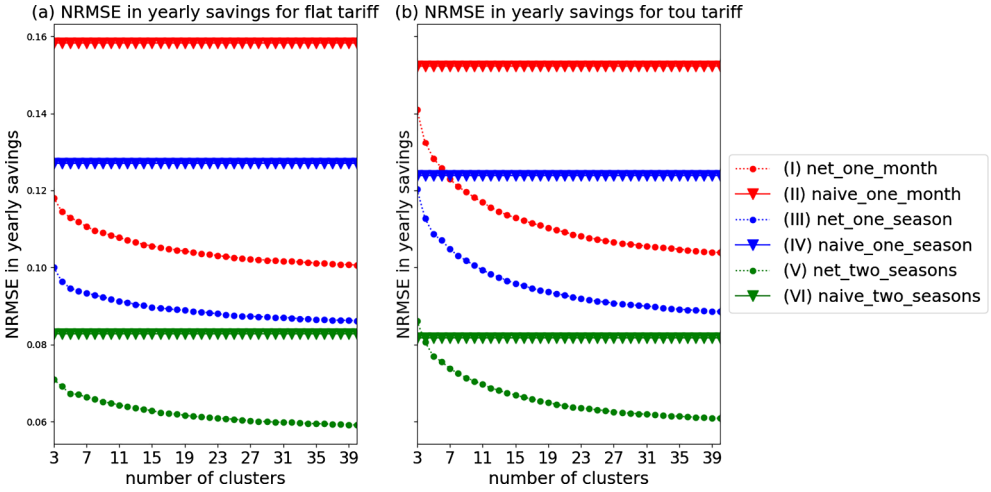

Fig. 9. Errors in estimated yearly savings under (a) a flat and (b) a ToU tariff when applying the proposed net meter clustering method with (I) one month, (III) one season & (V) two seasons of input data and the naive forecasting method with (II) one month, (IV) one season & (VI) two seasons of input data vs number of seasonal clusters per season.

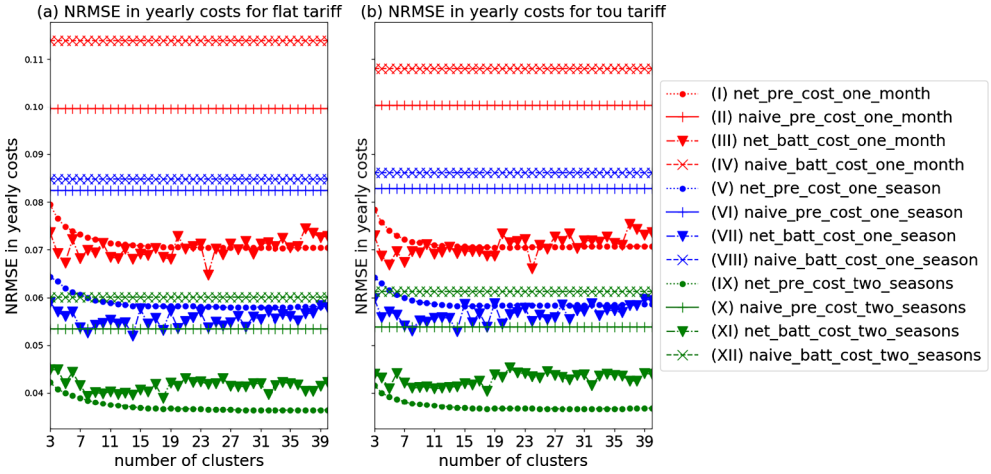

Fig. 10. Errors in estimated yearly costs before and after a battery is installed under (a) a flat and (b) a ToU tariff when applying the proposed net meter clustering method with (I, III) one month, (V, VII) one season & (IX, XI) two seasons of input data and the naive forecasting method with (II, IV) one month, (VI, VIII) one season & (X, XII) two seasons of input data vs number of seasonal clusters per season. (Note "pre_cost" and "batt_cost" respectively indicate yearly costs before and after a battery install.)

evaluated tariff structures and input data lengths. This indicates DBI can potentially be used as a metric for our end application. Hence, when a new dataset is provided, instead of going through different numbers of seasonal clusters and comparing the end results, the mean seasonal DBI values can potentially be used to directly select the best number of seasonal clusters which heavily reduces the computational costs.

4.3.2. Errors in NPVs and optimal sizes

NRMSEs in NPVs at the end of a battery's lifetime against the number of seasonal clusters for both naive forecasting case and net meter clustering case with different input data lengths are displayed in Fig. 12. Again the net meter clustering method has better performances compared to the naive forecasting approach for almost all tested scenarios except for one case where two-season input data and three

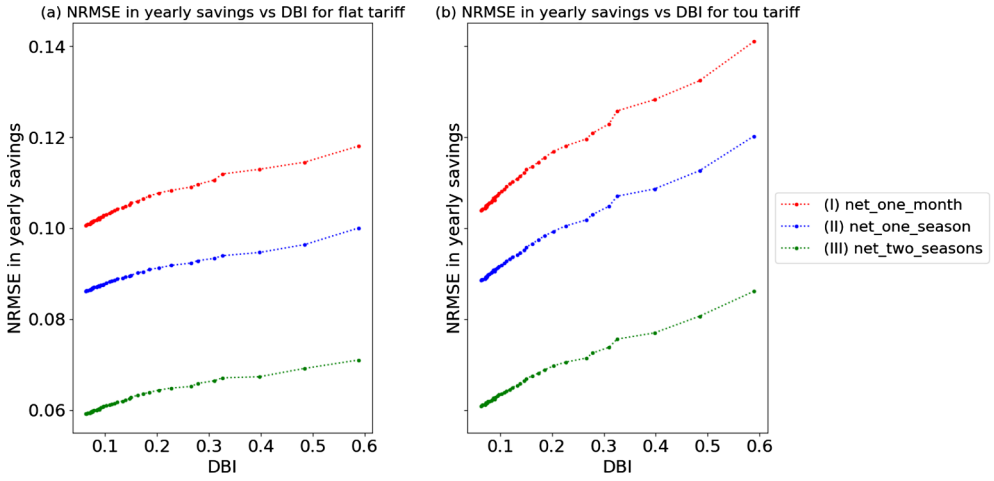

Fig. 11. Errors in estimated yearly savings under (a) a flat and (b) a ToU tariff when applying the proposed net meter clustering method with (I) one month, (II) one season & (III) two seasons of input data vs mean seasonal DBIs.

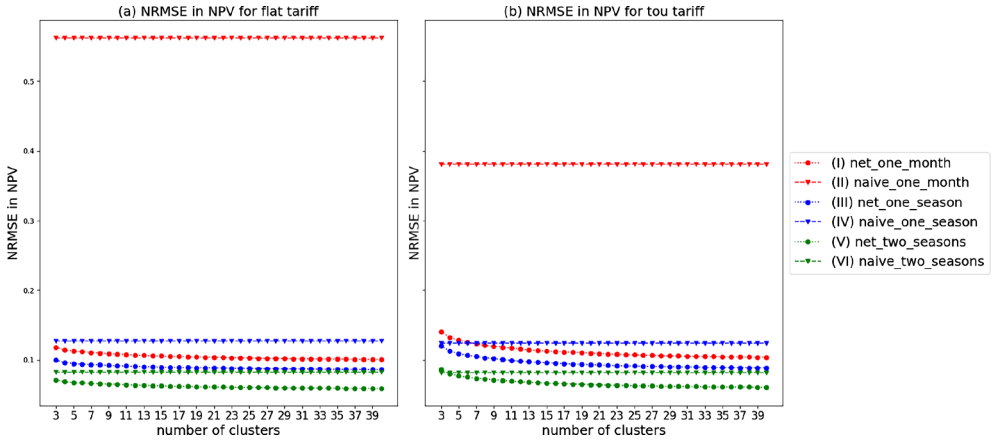

Fig. 12. Errors in estimated end NPV under (a) a flat and (b) a ToU tariff when applying the proposed net meter clustering method with (I) one month, (III) one season & (V) two seasons of input data and the naive forecasting method with (II) one month, (IV) one season & (VI) two seasons of input data vs number of seasonal clusters.

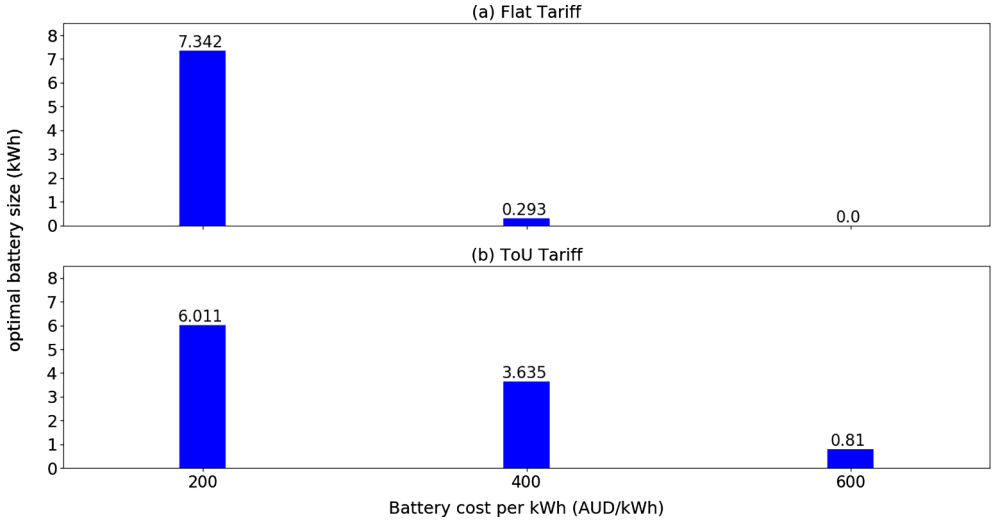

Fig. 13. Under (a) a flat tariff or (b) a ToU tariff, the mean optimal battery size derived using a full year of data.

seasonal clusters are applied. The differences in NRMSEs between the two methods are extremely large when only one month of data is used to extrapolate other data in a year. As a result, this shows the net meter clustering produces much better estimations on the profitability of installing a battery system compared to the naive forecasting model. This indicates that with a small amount of gross/net meter data, the net metering clustering approach is able to help the customers to have better ideas of whether they would make a profit or loss at the end of

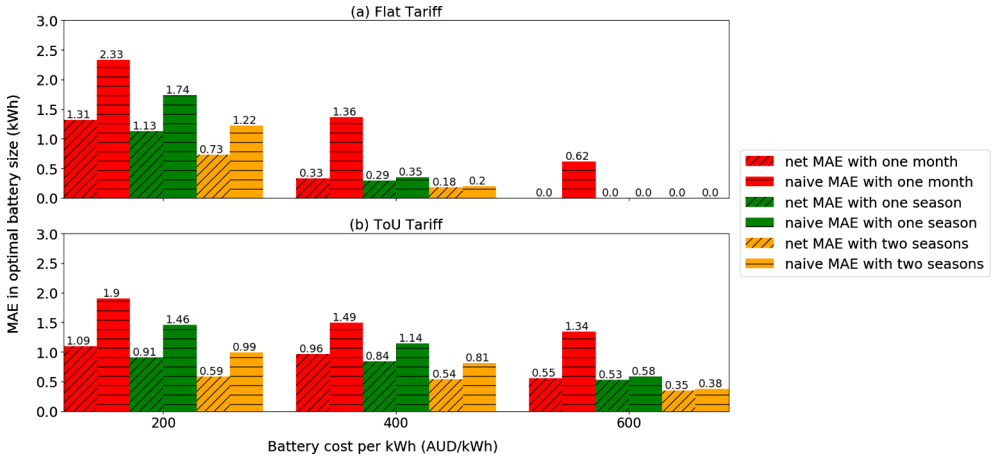

Fig. 14. Under (a) a flat tariff or (b) a ToU tariff, the MAE in estimated optimal battery sizes using the net meter clustering approach and naive forecasting method for different battery price ranges and input data lengths.

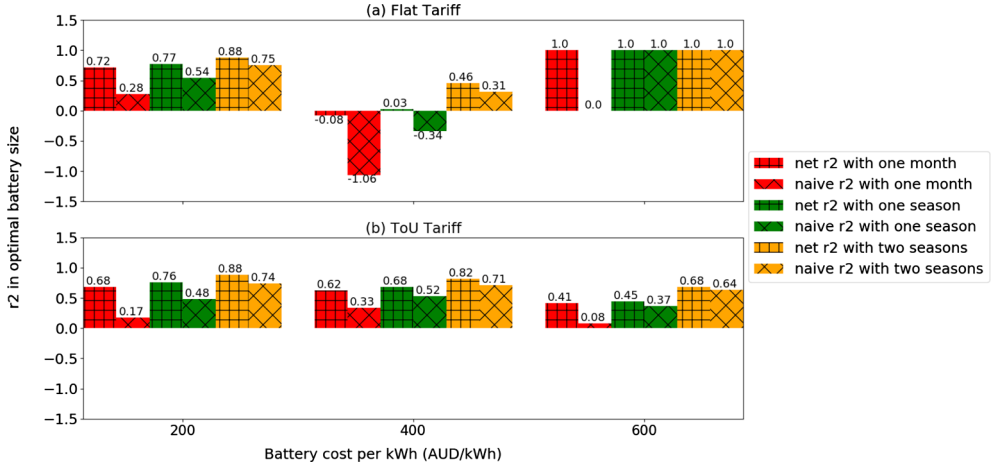

Fig. 15. Under (a) a flat tariff or (b) a ToU tariff, the mean R-squared value in estimated optimal battery sizes using the net meter clustering approach and naive forecasting method for different battery price ranges and input data lengths.

the battery lifetime.

Figs. 13–15 illustrate the mean true optimal battery size derived by the ideal case where all the data is provided, the mean absolute error (MAE) and r-squared value (R^2) in optimal battery sizes for the net meter approach and naive forecasting method under various battery price ranges. We also assume a constant installation price of $ 400. Both tariff structures (flat and ToU) are evaluated. We use 40 seasonal clusters for the net meter clustering approach and average the results for all the input data scenarios in Table 1. The net meter clustering model outperforms the naive forecasting method in terms of MAEs and r2 values for most battery and installation cost ranges, except for the cases where the true optimal sizes are quite close to zero. For the low battery price range ($200 per kWh), the developed model achieves r-squared values of 0.72 and 0.68 using a month of input data under the specified flat and ToU tariff, which is a quite good level of accuracy. At a lower cost range, both methods show better r2 values compared to medium battery costs. This is expected as the price increases, the optimal size tends to shift towards zero which means its variance will be much smaller compared to the residual sum of squares. Overall for the medium and large price ranges ($400-$600/kWh), the optimal battery sizes computed for ToU are larger compared to flat tariff. This means for these customers, ToU is a better option in terms of financial returns if they decide to install a battery as it will probably take a while for battery costs to drop to $200 per kWh.

Overall, by applying our proposed model with net meter energy data clustering on the test set of 262 Australian solar customers, we have obtained much better results in terms of estimated annual savings, costs before and after battery installations, end NPV and optimal sizes compared to the baseline naive forecasting approach. Furthermore, with a limited amount of net/gross meter energy data, the model still produces results with satisfactory accuracies.

For end-users who do not have easy access to enough historical smart meter data, the net metering clustering approach could be used to predict their annual electricity costs and battery profitability under different tariff structures. As a result, our proposed model could be implemented as a feature of a home energy recommendation tool to help residential customers make better tariff selection and battery purchase decisions with loose requirements on the length and quality of the input data. Moreover, installers and utilities who are likely to deal with customers with insufficient net meter data during the ongoing net meter rollouts, could utilise this technique as a recommendation service for their customers. They could also gain valuable insights on the impacts of tariff offers and battery prices on the electricity bills of their customers and make better predictions of the solar/battery market trends with a small amount of net/gross meter energy data.

5. Conclusion and future work

In this study, we perform a clustering analysis on net meter energy data and demonstrate that we could apply the correlations between seasonal cluster distributions to develop a battery sizing model that is quite robust to limited amount of input net meter energy data. As net

meters will eventually replace gross meters, we hope our approach could assist the techno-economic assessments of PV-integrated battery systems for households, installers and utilities, who often do not have sufficient historical generation and consumption data.

For future work, as we only use a single optimisation objective of maximising self-consumption, it would be interesting to see how well our approach could perform for other optimisation goals such as battery degradation reduction, peak demand reduction or price arbitrage.

The dataset we use is from solar customers in Australia so it would be worthwhile to apply our model to a dataset with customers in other countries to see how well our approach generalises in a different region.

The temporal resolution adopted in this study is half-hour net meter energy data, it could be interesting to explore what data granularity optimises the trade-offs between computations and performances of our proposed model.

Acknowledgements

The authors would like to thank Solar Analytics for providing the dataset for this study and J. Ma for some insightful suggestions that assisted this work. The authors acknowledge scholarships from the Research Training Program provided by the Australian Government.

References

[1] Australian Energy Council (AEC). Renewable energy in australia – how do we really compare?; 2016. https://www.energycouncil.com.au/media/1318/2016-06-23_aec-renewables-factsheet.pdf [Online; accessed 01-June-2018].
[2] Australian PV Institute (APVI). Australian pv institute (apvi) solar map, funded by the australian renewable energy agency; 2018. https://pv-map.apvi.org.au [Online; accessed 25-April -2018].
[3] Jäger-Waldau A. Pv status report 2017; 2017. http://publications.jrc.ec.europa.eu/repository/bitstream/JRC108105/kjna28817enn.pdf [Online; accessed 01-July-2018].
[4] Poruschi L, Ambrey CL, Smart JC. Revisiting feed-in tariffs in Australia: a review. Renew Sustain Energy Rev 2018;82(October 2016):260–70. https://doi.org/10.1016/j.rser.2017.09.027.
[5] Renewable Energy Policy for the 21st Century (REN21). Renewables 2017 global status report; 2017. http://www.ren21.net/gsr-2017/ [Online; accessed 15-June-2018].
[6] Ramarez FJ, Honrubia-Escribano A, Gamez-Lazaro E, Pham DT. Combining feed-in tariffs and net-metering schemes to balance development in adoption of photovoltaic energy: comparative economic assessment and policy implications for European countries. Energy Policy 2017;102:440–52.
[7] Smart Energy Council (SEC). Australian energy storage market analysis; 2018. https://www.smartenergy.org.au/sites/default/files/uploaded-content/field_f_content_file/australian_energy_storage_market_analysis_report_sep18_final.pdf [Online; accessed 15-September-2018].
[8] G.S.E. Solution, Grid-Connected PV Systems with Battery Storage; 2015.
[9] Linssen J, Stenzel P, Fleer J. Techno-economic analysis of photovoltaic battery systems and the influence of different consumer load profiles. Appl Energy 2017;185:2019–25. https://doi.org/10.1016/j.apenergy.2015.11.088.
[10] Motlagh O, Paevere P, Hong TS, Grozev G. Analysis of household electricity consumption behaviours: impact of domestic electricity generation. Appl Math Comput 2015;270:165–78. https://doi.org/10.1016/j.amc.2015.08.029.
[11] Chicco G, Ilie IS. Support vector clustering of electrical load pattern data. IEEE Trans Power Syst 2009;24(3):1619–28. https://doi.org/10.1109/TPWRS.2009.2023009.
[12] Stephen B, Mutanen AJ, Galloway S, Burt G, Jarventausta P. Enhanced load profiling for residential network customers. IEEE Trans Power Delivery 2014;29(1):88–96. https://doi.org/10.1109/TPWRD.2013.2287032.
[13] Hsiao YH. Household electricity demand forecast based on context information and user daily schedule analysis from meter data. IEEE Trans Industr Inf 2015;11(1):33–43. https://doi.org/10.1109/TII.2014.2363584.
[14] Yildiz B, Bilbao JI, Dore J, Sproul A. Household electricity load forecasting using historical smart meter data with clustering and classification techniques. IEEE PES ISGT Asia 2018:873–9. https://doi.org/10.1109/ISGT-Asia. 2018.8467837.
[15] Yang Y, Bremner S, Menictas C, Kay M. Battery energy storage system size determination in renewable energy systems: a review. Renew Sustain Energy Rev 2018;91(January):109–25. https://doi.org/10.1016/j.rser.2018.03.047.
[16] Khalilpour R, Vassallo A. Planning and operation scheduling of PV-battery systems: a novel methodology. Renew Sustain Energy Rev 2016;53:194–208. https://doi.org/10.1016/j.rser.2015.08.015.
[17] Astaneh M, Roshandel R, Dufo-López R, Bernal-Agustín JL. A novel framework for optimization of size and control strategy of lithium-ion battery based off-grid renewable energy systems. Energy Convers Manage 2018;175(July):99–111. https://doi.org/10.1016/j.enconman.2018.08.107.
[18] Rodríguez-Gallegos CD, Yang D, Gandhi O, Bieri M, Reindl T, Panda SK. A multi-objective and robust optimization approach for sizing and placement of PV and batteries in off-grid systems fully operated by diesel generators: an Indonesian case study. Energy 2018;160:410–29. https://doi.org/10.1016/j.energy.2018.06.185.
[19] Schram WL, Lampropoulos I, van Sark WG. Photovoltaic systems coupled with batteries that are optimally sized for household self-consumption: assessment of peak shaving potential. Appl Energy 2018;223(April):69–81. https://doi.org/10.1016/j.apenergy.2018.04.023.
[20] Talent O, Du H. Optimal sizing and energy scheduling of photovoltaic-battery systems under different tariff structures. Renew Energy 2018;129:513–26. https://doi.org/10.1016/j.renene.2018.06.016.
[21] Berrueta A, Heck M, Jantsch M, Ursúa A, Sanchis P. Combined dynamic programming and region-elimination technique algorithm for optimal sizing and management of lithium-ion batteries for photovoltaic plants. Appl Energy 2018;228(February):1–11. https://doi.org/10.1016/j.apenergy.2018.06.060.
[22] Pflaum P, Alamir M, Lamoudi MY. Battery sizing for PV power plants under regulations using randomized algorithms. Renew Energy 2017;113:596–607. https://doi.org/10.1016/j.renene.2017.05.091.
[23] Talavera DL, Muñoz-Rodriguez FJ, Jimenez-Castillo G, Rus-Casas C. A new approach to sizing the photovoltaic generator in self-consumption systems based on cost-competitiveness, maximizing direct self-consumption. Renew Energy 2019;130:1021–35. https://doi.org/10.1016/j.renene.2018.06.088.
[24] Ali A, Mohd Nor N, Ibrahim T, Fakhizan Romlie M. Sizing and placement of battery-coupled distributed photovoltaic generations. J Renew Sustain Energy 2017;9(5). https://doi.org/10.1063/1.4995531.
[25] Aghamohammadi MR, Abdolahinia H. A new approach for optimal sizing of battery energy storage system for primary frequency control of islanded Microgrid. Int J Electr Power Energy Syst 2014;54:325–33. https://doi.org/10.1016/j.ijepes.2013.07.005.
[26] Jannesar MR, Sedighi A, Savaghebi M, Guerrero JM. Optimal placement, sizing, and daily charge/discharge of battery energy storage in low voltage distribution network with high photovoltaic penetration. Appl Energy 2018;226(March):957–66. https://doi.org/10.1016/j.apenergy.2018.06.036.
[27] Ru Y, Kleissl J, Martinez S. Storage size determination for grid-connected photovoltaic systems. IEEE Tran Sustain Energy 2013;4(1):68–81. https://doi.org/10.1109/TSTE.2012.2199339. arXiv: 1109.4102.
[28] Weniger J, Tjaden T, Quaschning V. Sizing of residential PV battery systems. Energy Procedia 2014;46:78–87. https://doi.org/10.1016/j.egypro.2014.01.160.
[29] Hemmati R, Saboori H. Stochastic optimal battery storage sizing and scheduling in home energy management systems equipped with solar photovoltaic panels. Energy Build 2017;152:290–300. https://doi.org/10.1016/j.enbuild.2017.07.043.
[30] Khalilpour KR, Vassallo A. Technoeconomic parametric analysis of PV-battery systems. Renew Energy 2016;97:757–68. https://doi.org/10.1016/j.renene.2016.06.010.
[31] Quoilin S, Kavvadias K, Mercier A, Pappone I, Zucker A. Quantifying self-consumption linked to solar home battery systems: Statistical analysis and economic assessment. Appl Energy 2016;182:58–67. https://doi.org/10.1016/j.apenergy.2016.08.077.
[32] Schopfer S, Tiefenbeck V, Staake T. Economic assessment of photovoltaic battery systems based on household load profiles. Appl Energy 2018;223(November 2017):229–48. https://doi.org/10.1016/j.apenergy.2018.03.185.
[33] Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 2012;42(1):68–80. https://doi.org/10.1016/j.energy.2011.12.031.
[34] Yildiz B, Bilbao J, Dore J, Sproul A. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. Appl Energy 2017. https://doi.org/10.1016/j.apenergy.2017.10.014. http://linkinghub.elsevier.com/retrieve/pii/S0306261917314265.
[35] Costa N, Matos I. Inferring daily routines from electricity meter data. Energy Build 2016;110:294–301. https://doi.org/10.1016/j.enbuild.2015.11.015.
[36] Piao M, Ryu KH. Local characterization-based load shape factor definition for electricity customer classification. IEEJ Trans Electr Electron Eng 2017;12:S110–6. https://doi.org/10.1002/tee.22424.
[37] Viegas JL, Vieira SM, Melício R, Mendes VM, Sousa JM. Classification of new electricity customers based on surveys and smart metering data. Energy 2016;107:804–17. https://doi.org/10.1016/j.energy.2016.04.065.
[38] Abreu JM, Câmara Pereira F, Ferrão P. Using pattern recognition to identify habitual behavior in residential electricity consumption. Energy Build 2012;49:479–87. https://doi.org/10.1016/j.enbuild.2012.02.044.
[39] Kwac J, Flora J, Rajagopal R. Household energy consumption segmentation using hourly data. IEEE Trans Smart Grid 2014;5(1):420–30. https://doi.org/10.1109/TSG.2013.2278477.
[40] Tsekouras GJ, Hatziargyriou ND, Dialynas EN. Two-stage pattern recognition of load curves for classification of electricity customers. IEEE Trans Power Syst 2007;22(3):1120–8. https://doi.org/10.1109/TPWRS.2007.901287.
[41] Chicco G, Napoli R, Piglione F. Comparisons among clustering techniques for electricity customer classification. IEEE Trans Power Syst 2006;21(2):933–40. https://doi.org/10.1109/TPWRS.2006.873122.
[42] McLoughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl Energy 2015;141:190–9. https://doi.org/10.1016/j.apenergy.2014.12.039.
[43] Gerbec D, Gašperič S, Šmon I, Gubina F. Allocation of the load profiles to consumers using probabilistic neural networks. IEEE Trans Power Syst 2005;20(2):548–55. https://doi.org/10.1109/TPWRS.2005.846236.
[44] Chicco G, Napoli R, Pigline F. Comparisons among clustering techniques for electricity customer classification. IEEE Trans Power Syst 2006;21(2):1–7. https://doi.org/10.1109/TPWRS.2006.873122.

[45] Florita AR, Brackney LJ, Otanicar TP, Robertson J. Classification of commercial building electrical demand profiles for energy storage applications. J Sol Energy Eng 2013;135(3):031020. https://doi.org/10.1115/1.4024029http://solarenergyengineering.asmedigitalcollection.asme.org/article.aspx?doi=10.1115/1.4024029.
[46] Solar Analytics. Connect with your solar; 2018. https://www.solaranalytics.com/au/ [Online; accessed 02-August-2018].
[47] Wattwatchers. Wattwatchers: Super-smart devices for energy monitoring; 2018. https://wattwatchers.com.au/ [Online; accessed 02-November-2018].
[48] Haben S, Singleton C, Grindrod P. Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Trans Smart Grid 2016;7(1):136–44. https://doi.org/10.1109/TSG.2015.2409786.
[49] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. Appl Energy 2014;135:461–71. https://doi.org/10.1016/j.apenergy.2014.08.111.
[50] Hino H, Shen H, Murata N, Wakao S, Hayashi Y. A versatile clustering method for electricity consumption pattern analysis in households. IEEE Trans Smart Grid 2013;4(2):1048–57. https://doi.org/10.1109/TSG.2013.2240319. http://ieeexplore.ieee.org/ielx7/5165411/6517533/06484217.pdf?tp=&arnumber=6484217&isnumber=6517533%5Cnhttp://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6484217.
[51] Bureau of Meteorology. Renewable energy in australia – how do we really compare? http://www.bom.gov.au/climate/glossary/seasons.shtml [Online; accessed 19-June-2018] (no date).
[52] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth berkeley symposium on mathematical statistics and probability: statistics, vol. 1. Berkeley, Calif.: University of California Press; 1967. p. 281–97. https://projecteuclid.org/euclid.bsmsp/1200512992.
[53] Bottou L, Bengio Y. Convergence properties of the k-means algorithms. Adv Neural Inform Process Syst 1995:585–92.
[54] Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1979(2):224–7.
[55] Anderson TW, Anderson TW, Anderson TW, Anderson TW, Mathématicien E-U. An introduction to multivariate statistical analysis vol. 2. New York: Wiley; 1958.
[56] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[57] Ho TK. Random decision forests. Document analysis and recognition, 1995, proceedings of the third international conference on, vol. 1. IEEE; 1995. p. 278–82.
[58] Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–40.
[59] Géron A. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media Inc; 2017.
[60] Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B (Methodol) 1996:267–88.
[61] Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Softw 2010;36(11):1–13.
[62] Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Briefings Bioinformatics 2017. https://doi.org/10.1093/bib/bbx124.
[63] Daniel H. boruta py; 2016. https://github.com/scikit-learn-contrib/boruta_py.
[64] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13:281–305http://dl.acm.org/citation.cfm?id=2188385.2188395.
[65] Abdulla K, Steer K, Wirth A, de Hoog J, Halgamuge S. The importance of temporal resolution in evaluating residential energy storage. Power & energy society general meeting, 2017 IEEE. IEEE; 2017. p. 1–5.
[66] Tang R, Abdulla K, Leong PH, Vassallo A, Dore J. Impacts of temporal resolution and system efficiency on PV battery system optimisation. In: 2017 Asia-Pacific solar research conference. http://apvi.org.au/solar-research-conference/wp-content/uploads/2017/12/029_R-Tang_DI_Paper_Peer-reviewed.pdf.
[67] Solar Choice. Is home solar battery storage worth it? (jan 2018 update); 2018. https://www.solarchoice.net.au/blog/home-solar-battery-storage-worth-it-2018 [Online; accessed 29-August-2018].
[68] Solar Quotes. Solar battery storage comparison table; 2018. https://www.solarquotes.com.au/battery-storage/comparison-table/ [Online; accessed 16-September-2018].