An Unsupervised Data-driven Approach for Behind-the-Meter Photovoltaic Power Generation Disaggregation

Keda Pan[a], Zhaohua Chen[b], Chun Sing Lai[b,c], Changhong Xie[b], Dongxiao Wang[b,d], Xuecong Li[b], Zhuoli Zhao[b], Ning Tong[b], Loi Lei Lai[a,b]

[a] Department of Control Engineering, School of Automation, Guangdong University of Technology, Guangzhou 510006, China

[b] Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou 510006, China

[c] Brunel Interdisciplinary Power Systems Research Centre, Department of Electronic and Electrical Engineering, Brunel University London, London UB8 3PH, UK

[d] System Design and Engineering Department, Australia Energy Market Operator, Melbourne 3000, Australia

Corresponding author: l.l.lai@ieee.org (L. L. Lai)

Abstract: An increasing number of behind-the-meter (BtM) rooftop photovoltaic (PV) panels is being installed and maintained by site owners. However, invisible PV power generation (PVPG) will lead to the difficulty for system operators in power dispatch and affect the safety and stability of the power system. To better quantify BtM PVPG, a novel unsupervised data-driven disaggregation method freedom from PV system physical model assumption for BtM PVPG is proposed. After clustering the prosumers' net load curves, a PVPG sensitivity estimation model is firstly built, based on the net load with approximate energy consumption (EC) and the corresponding irradiation data obtained from the pairing date. Then, an EC sensitivity model is developed according to the net load and temperature of the date with similar irradiation. Finally, a new net load disaggregation model is constructed by the PVPG sensitivity model with EC compensation. Case study based on Ausgrid data shows that the proposed method provides a better quality BtM PVPG disaggregation. The disaggregation accuracy improves by 5.06%-5.87% as compared to the state-of-the-art methods.

Keywords: Behind-the-meter, net load disaggregation, energy consumption, PV power generation, net load, data-driven

Abbreviations and notations

Abbreviations

AEMO	Australian Energy Market Operator
BtM	Behind-the-meter
DBI	Davies-Bouldin index
DER	Distributed energy resources
DHI	Diffuse horizontal irradiation
DNI	Direct normal irradiation
DNN	Deep neural network
DTW	Dynamic time warping
EC	Energy consumption
ED	Euclidean distance
FCM	Fuzzy C-means
GMM	Gaussian mixture models
KNN	*k*-nearest neighbor
LSTM	Long short-term memory
MAPE	Mean absolute percentage error
MLP	Multilayer perceptron
NEM	National electricity market
PV	Photovoltaic
PVPG	Photovoltaic power generation
RMSE	Root mean square error
SVR	Support vector regression
DWT	Discrete wavelet transform

Notations

c	Number of clustering centers $\mathbf{D}_{\mathrm{C}}$
	Matched EC day set
$\mathbf{D}_{\mathrm{G}}$	Matched PVPG day set
f	Time resolution of data set
$f_{\mathrm{I,G}}(\bullet)$	Fitting function between irradiation and PVPG
$f'_{\mathrm{I,G}}(\bullet)$	Compensatory fitting function between irradiation and PVPG
$f_{\mathrm{T,C}}(\bullet)$	Fitting function between ambient temperature and EC
I_{DHI}	DHI
I_{DNI}	DNI
$k1$	Number of nearest neighbors of PVPG sensitivity model
$k2$	Number of nearest neighbors of EC sensitivity model
P'_{C}	Disaggregated EC
P'_{G}	Disaggregated PVPG
$\mathbf{P}_{\mathrm{C}}$	Set of EC P_{C}
$\mathbf{P}_{\mathrm{G}}$	Set of PVPG P_{G}
$\mathbf{P}_{\mathrm{N}}$	Set of net load P_{G}
t_{e}	Starting point of evening
t_{s}	Starting point of day time
$\mathbf{t}$	Set of a whole day time step t
$\mathbf{t}_{\mathrm{day}}$	Set of day time step t_{day}
$\mathbf{t}_{\mathrm{middle}}$	Set of the middle part of $\mathbf{t}_{\mathrm{day}}$
$\mathbf{t}_{\mathrm{night}}$	Set of evening time step t_{night}
$\mathbf{T}$	Set of ambient temperature T
$\Delta\mathbf{I}_{\mathrm{DHI}}$	Set of DHI difference ΔI_{DHI}
$\Delta\mathbf{I}_{\mathrm{DNI}}$	Set of DNI difference ΔI_{DNI}
$\Delta\mathbf{P}_{\mathrm{C}}$	Set of EC difference ΔP_{C}
$\Delta\tilde{\mathbf{P}}_{\mathrm{C}}$	Set of estimated/compensatory EC difference
$\Delta\mathbf{P}_{\mathrm{G}}$	Set of PVPG difference ΔP_{G}

$\Delta\hat{\mathbf{P}}_{\mathrm{G}}$	Set of estimated value of PVPG difference $\Delta\tilde{P}_{\mathrm{G}}$
$\Delta\mathbf{P}_{\mathrm{N}}$	Set of net load difference ΔP_{N}
$\Delta t_{\mathrm{middle}}$	Duration time of $\mathbf{t}_{\mathrm{middle}}$
$\Delta\mathbf{T}$	Set of ambient difference ΔT

1. Introduction

The greenhouse effect caused by excessive carbon emissions has led to the deterioration of the natural environment and frequent extreme weather. In order to achieve decarbonization, countries are striving to get rid of dependence on fossil energy and committed to a carbon neutrality goal [1]. Due to the technological advancement, under several conditions, the cost of renewables sources like photovoltaic (PV) can now be comparable to fossil fuels [2], [3]. Distributed PV power generation (PVPG) can effectively reduce greenhouse gas emissions [4] and there is an increasing interest in investing in the technology. By 2040, the distributed generation capacity is expected to double or even triple, and Australian Energy Market Operator (AEMO) modelling projects distributed energy resources (DER) could provide 13% to 22% of total underlying annual national electricity market (NEM) energy consumption (EC) [5]. However, considering the economy, timeliness of information and customer privacy, most behind-the-meter (BtM) rooftop distributed PV systems do not have complete or up-to-date information directly provided to the grids [6], resulting in invisible or inaccurate recorded PVPG data. Although the net load can reflect the behaviors of prosumers to some extent, the comprehensive expression of the offset part of the PVPG and EC is still not discussed. This obliterated data will cause problems such as the difficulties in power scheduling, disabled designed reverse power flow [7] and mismatching the energy storage capacity between solar availability and electricity demand [8]. Therefore, to allow DER to maximize their contribution and potential to the power system, it is indispensable to develop an effective BtM disaggregation method.

Based on the metered power information for modelling, BtM PVPG disaggregation is generally classified as 1) supervised disaggregation methods using fully separated metered data of customers to be disaggregated, 2) semi-supervised disaggregation methods using the partial metered data of PV system or the recording data of PV proxy and 3) unsupervised disaggregation methods only use metered net load data. Considering the application environment, only semi-supervised and unsupervised disaggregation methods are studied in this paper.

As the EC data and PVPG data are not available, a common BtM semi-supervised disaggregation idea is adopted to map the relationship between net load and PVPG information by setting representative PV system proxy, and extend this mapping relationship to the target PV sites.

In [9], by using data dimension reduction methodology, a small number of representative solar sites is selected according to a four-month collected PVPG data of all sites. Consequently, the BtM PVPG value of all sites is estimated by the PVPG data of representative solar sites and external variables. However, in the BtM scenario, it is unrealistic to obtain the data of all PV sites for a period of time. With the set of fully observable EC and PVPG of partial customers in [10], the net load customers under the same lateral secondary distribution transformer are disaggregated by employing a designed semi-supervised signal separation optimization algorithm. Assuming that a small number of distributed PV systems are available, support vector regression (SVR) is implemented in [11] to build a capacity estimation model based on the constructed net load feature, and calculates the disaggregated PV output. Whereas the known distributed PV systems need to have the same system geometry of other distributed PV systems in the area. Unlike obtaining the metered PVPG proxy data, in [6] by a given PV panels size, tile and azimuth of some users, the parameter estimation model is established through deep neural network (DNN). Then the disaggregated PVPG is calculated by the physics-based model according to the estimated parameters of the PV panels. In [12], the net load disaggregation model is constructed based on the accessible PV system data. By assuming that users of a building have consistency electricity demand behaviors before and after installation of PV system, disaggregated EC can be obtained by the value of the PV system post-installed building consumption at the comparable timestamp. However, the model will fail if similar weather conditions do not occur. Besides, it can almost be considered as a supervised disaggregation method, because the separated EC data has been known before the PV system is installed.

The application of the above-mentioned proxy based semi-supervised disaggregation methods is impractical. Because of the difficulties of quantifying and equating the PV system geometry and physical characteristics of multiple target PV systems and setting a proxy to match them. Even if the matching proxy already exists, data collection time is still required, which will delay the implementation of the disaggregation plan and increase time costs.

To avoid setting up a PV proxy, unsupervised disaggregation methods that only use net load data are designed. By utilizing PV system physical model assumption, parameters of PV panels are estimated through swarm intelligence algorithms in [13], while PVPG is obtained with the correction of modeling the temperature effect. But the method relies too much on the period when the building is idle, which will greatly reduce the scope of applications. Depending on the PV system geometry, the equivalent capacity is firstly estimated by employing maximal information coefficient-based grid search [14]. The net load is iteratively disaggregated into three parts by minimizing the correlation between the disaggregated PVPG and EC.

Although the unsupervised disaggregation methods proposed in [13] and [14] only use net load data and can effectively improve the usage conditions, the methods rely on a specific PV system physical model assumptions. The disaggregation results will be inaccurate if the PV arrays are degraded or the PV physical model is incorrectly assumed [2].

To solve the problem of overdependence on assumption of PV system physical model and PV system geometry, a BtM PVPG disaggregation method based on nonlinear programming is proposed in [15]. The assumption of the PV system physical model is only used for the conversion of output efficiency affected by the temperature of the PV panel. However, the model does not consider the influence of temperature on EC behavior, resulting in errors in the disaggregation results.

Table 1 presents an overview of the recent works in BtM PVPG disaggregation. It can be observed that, although compared to semi-supervised disaggregation methods, unsupervised disaggregation methods have their unique advantages, the existing unsupervised disaggregation methods depend on the setting of the physical model, which will bring this unsupervised disaggregation method to another error dilemma.

Table 1
Comparison of recent BtM disaggregation studies.

Work	Disaggregation type	Disaggregation framework	Data needed	Algorithms needed	Disaggregation level
Shaker et al. [9]	Semi-supervised	Data-driven	Net load, PV site locations, meteorological, PVPG of proxy and a 4-month separated metered data of all PV sites.	K-means, principal component analysis, linear regression, Kalman filter, multilayer perceptron, and wavelet neural network.	Aggregated net load of 405 customers and PV sites.
Bu et al. [10]	Semi-supervised	Data-driven	Net load, separated metered data of proxy.	Spectral clustering and designed semi-supervised signal separation optimization.	Aggregated net load of 1120 residential customers and 337 PV sites.
Li et al. [11]	Semi-supervised	Data-driven	Net load, PVPG and capacity of proxy and meteorological.	Support vector regression.	Net load of a single residential house.
Mason et al. [6]	Semi-supervised	Data-driven with PV system geometry	Net load, PV system physical model parameters.	Linear regression and deep neural network.	Aggregated net load of 1000 residential customers and PV sites.
Stainsby et al. [12]	Almost fully supervised	Data-driven	Net load, EC before PV system installed, date of PV system installed and meteorological.	Pairing, comparison and displacement.	Net load of a single residential house or building.
Chen et al. [13]	Unsupervised	PV system geometry	Net load, PV sites location and meteorological.	Swarm intelligence algorithms.	Net load of a single building.
Wang et al. [14]	Unsupervised	Data-driven combined with PV system geometry	Net load and meteorological, PV system physical model parameters.	Grid search and maximal information coefficient.	Zonal level.
Sossan et al. [15]	Unsupervised	Data-driven combined with PV system geometry	Net load and meteorological, PV system physical model parameters.	Nonlinear programming.	Net load of a single house.
This work	Unsupervised	Data-driven	Net load and meteorological.	Clustering, *k*-nearest neighbor, multilayer perceptron and long short-term memory network.	Aggregated net load from 29 to 77 residential customers and PV sites.

As shown from the literature review, the semi-supervised methods based on the PV proxy mostly are data-driven with using machine learning-based principles for PVPG disaggregation, but the learning targets for training disaggregation models are from the additional installation PV proxy, which is difficult to ensure that its PVPG output characteristics are consistent with the target PV devices. While the unsupervised methods analyze the existing net load and perform disaggregation in combination with the assumption of PV system physical model, which avoids the problem of inconsistent output characteristics of semi-supervised methods based on PV proxy, but the incorrect assumptions of the PV system physical model are prone to introduce mismatched PVPG characteristics and resulting in decreased disaggregation accuracy.

In the existing work, the supervised and semi-supervised net load disaggregation methods have developed rapidly, while the unsupervised method has few breakthroughs, and there is no PV system physical model independent unsupervised disaggregation model. Therefore, in order to solve the restriction of the PV proxy and ease the dependence on the PV system physical model and geometry, an unsupervised data-driven net load disaggregation method only needs net load and meteorological data as input is proposed in this paper to close the research gap and focus on the practical application. The distinguished contributions of this paper are as follows:

- To the best of authors’ knowledge, this is the first unsupervised data-driven net load disaggregation method that completely omits PV physical model assumption and PV system geometry, with the consideration of PV conversion efficiency due to ambient temperature variation.
- The proposed unsupervised net load disaggregation method extracts the PVPG features by net load to build a BtM disaggregation model using machine learning algorithms. By avoiding the inconsistency of output characteristics caused by semi-supervised methods, PVPG is obtained by setting up proxy used as learning targets.
- A data-driven EC sensitivity model is proposed to refine the initial PVPG disaggregation model with meteorological data, which can effectively reflect the variations of net load caused by the energy inrush.

- The testing results show that the proposed unsupervised data-driven net load disaggregation methods have higher disaggregation accuracy and better disaggregation stability than the state-of-the-art unsupervised disaggregation methods.

2. Problem statement and data description

2.1. Problem statement

Distributed PV systems produce positive power when there is irradiation. For prosumers whose energy production is only generated by PVPG, the net load at night is equivalent to the EC. For distinction, the sets of daytime $\mathbf{t}_{\text{day}}$, evening $\mathbf{t}_{\text{night}}$ and a whole day $\mathbf{t}$ are written as follows:

$$\begin{cases} \mathbf{t}_{\text{day}} = \left\{ t_{\text{day}} \middle| t_{\text{day}} \in [t_{\text{s}}, t_{\text{s}}+1, \ldots, t_{\text{e}}-1] \right\} \\ \mathbf{t}_{\text{night}} = \left\{ t_{\text{night}} \middle| t_{\text{night}} \in [1,2,\ldots,t_{\text{s}}-1] \cup [t_{\text{e}}, t_{\text{e}}+1, \ldots, f] \right\} \end{cases} \tag{1}$$

subject to

$$\mathbf{t} = \mathbf{t}_{\text{day}} \cup \mathbf{t}_{\text{night}} = \left\{ t \middle| t \in [1,2,\ldots,f] \right\} \tag{2}$$

where t_{day}, t_{night} and t represent the time step of $\mathbf{t}_{\text{day}}$, $\mathbf{t}_{\text{night}}$ and, $\mathbf{t}$ respectively, t_{s} and t_{e} represent the starting point of daytime and evening respectively, f denotes the time resolution of the data set. Net load is defined with Eq. (3):

$$\mathbf{P}_{\text{N}} = \mathbf{P}_{\text{C}} - \mathbf{P}_{\text{G}} \tag{3}$$

where $\mathbf{P}_{\text{N}}$, $\mathbf{P}_{\text{C}}$ and $\mathbf{P}_{\text{G}}$ are the set of net load P_{N}, EC P_{C} and PVPG P_{G}, respectively. The time division and net load composition are shown in Fig. 1.

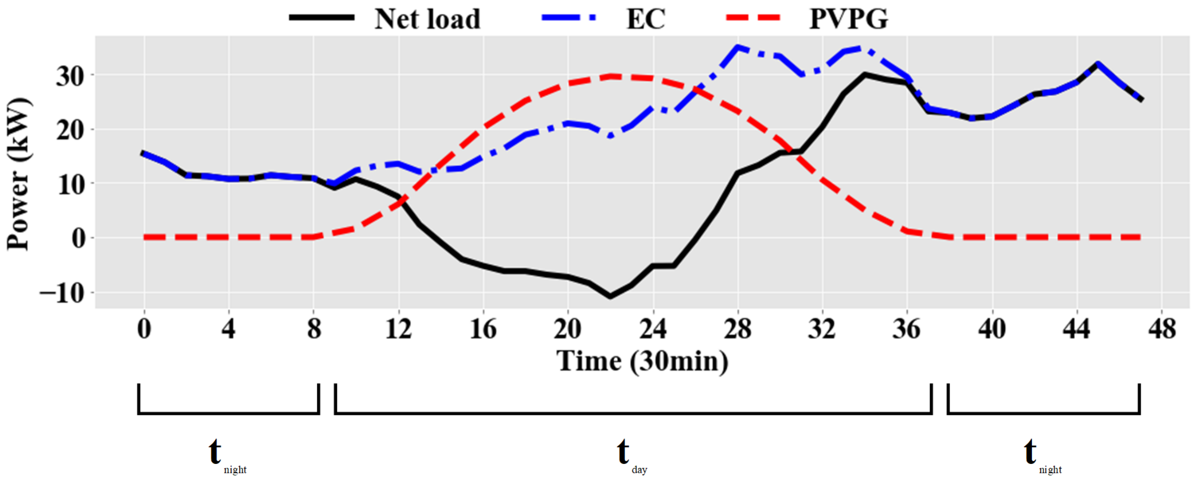


Fig. 1 Time division and net load composition

In most BtM situations, only the net load represented by the black line in Fig. 1 can be directly obtained, while the EC and PVPG corresponding to the blue and red lines respectively are unknown. The key problem to be solved in this paper is to calculate the values of EC and PVPG when only the net load and the corresponding meteorological data are available.

The meteorology time series is composed of ambient temperature T , direct normal irradiance (DNI) and diffuse horizontal irradiation (DHI) with a daily resolution of 24. To unify the format of net load data, average interpolation is applied to the raw meteorology data.

2.2. Data description

To verify the effectiveness of the proposed BtM PVPG disaggregation method, a public dataset incorporating semi-synthetic data is used. This dataset consists of 300 un-identified customers with only zip code recorded in Sydney and surrounding regional areas, with EC and rooftop PVPG separately measured [16]. The time span of the data set is from July 1, 2010 to June 30, 2013, with a resolution of 48 points per day. For Australia, as one of the countries with most serious extreme weathers, heatwave will lead to extreme peaks in demand, therefore a total of 17 heatwave days were marked and will be discussed specifically in the case study [17]. According to [18], 18 customers with anomalous general load data and controllable data were removed, and 13 customers in remote areas were further removed according to the area where the zip code belongs. The remaining 269 customers are further divided into 5 sub-regions based on their spatial distance. Considering that the data provided by the Ausgrid did not attach corresponding meteorological information, two methods were implemented to deal with this situation. 1) The raw PVPG data was removed and simulated by the model proposed in [19] based on the meteorology data from [20], the reality penetration level and distributed system capacity owned by each customer. This data set is mainly concerned with the accuracy of the proposed disaggregation method under the exclusion of the measurement errors caused by the mismatch between observed meteorological and PVPG data. 2) The raw PVPG data was used in conjunction with the meteorological data from the central area in the sub-regions [20]. This dataset will test the robustness of the disaggregation method using heterogeneous collection meteorological data of proximity site. Fig. 2 illustrates the distribution of customers in 5 sub-regions built by Google Earth [21]. The high latitude to bottom latitude represents the Sub-region 1 to Sub-region 5 respectively.

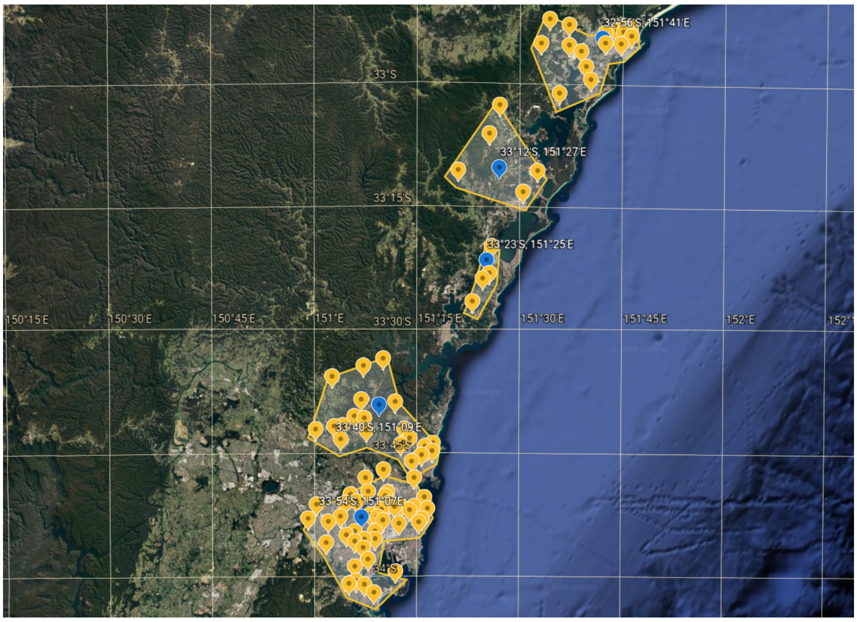


Fig. 2. Geographical location and area division of 269 customers

The yellow dots in Fig. 2 represent the location of the customers' zip codes, the closed yellow polylines represent the boundaries of the sub-regions, and the blue dots represent the meteorological collection points from the central area in the sub-regions.

The number of customers, generation capacity, original and designed PV penetration level [22] in each area are shown in Table 2. More detailed parameters for data description are added in Appendix A to facilitate reproducing of results.

Table 2
Parameters of the 5 sub-regions.

Sub-region	Number of Users	Daily Peak Load (kW)	Total Generation Capacity (kW)	Reality Penetration Level (%)	Designed Penetration Level (%)
1	77	87.62	119.77	32.23	33.73
2	63	71.27	104.45	30.53	32.17
3	29	34.91	47.97	36.42	35.74
4	44	50.57	63	40.61	38.51
5	56	57.58	107.53	44.96	43.70

The remainder of this paper is organized as follows. Section 2 describes the problem statement and experimental data. Section 3 proposes the framework and details of the disaggregation model. Section 4 introduces the evaluation metrics and method comparisons. Section 5 shows the effectiveness of the proposed model through the experimental results of the case study. Section 6 gives the conclusion.

3. Framework and proposed methodology

3.1. Framework

The basic idea is to refine the regional aggregated load according to the user's electricity curve shape by clustering, then the net load disaggregation with EC difference compensation is implemented for each cluster to decouple the PVPG curves. To achieve the above process, three stages called PVPG sensitivity modelling, EC sensitivity modelling and net load disaggregation stages are designed as follows:

1) PVPG sensitivity modelling stage

Due to the BtM environment, it is impossible to directly obtain integrated time series containing only PVPG from utility grid side, but by matching the date with approximate EC behavior, the PVPG difference can be obtained through the corresponding paired net load difference. Considering the strong correlation between PVPG and irradiation, a neural network based PVPG sensitivity model is built according to the irradiation difference and PVPG difference approximated by the related net load difference.

2) EC sensitivity modelling stage

In consistency with the data series of PVPG, series containing only EC information can only be obtained indirectly by the net load difference of pairing dates with approximated irradiation. Considering the strong drive of ambient temperature to EC behaviors, a neural network-based EC sensitivity model is built according to the ambient temperature difference and EC difference approximated by the related net load difference.

3) Net load disaggregation stage

In this stage, the PVPG sensitivity model is firstly amended by the output of EC sensitivity model and then innovatively transformed into BtM PVPG disaggregation model by utilizing the output characteristics of PV panels, the power output is to 0 when there is no radiation. Consequently, a more accurate net load disaggregation model including EC compensation is built.

The framework of the proposed BtM net load disaggregation methodology is shown in Fig. 3. The details of the three stages will be explained in subsequent sections later.

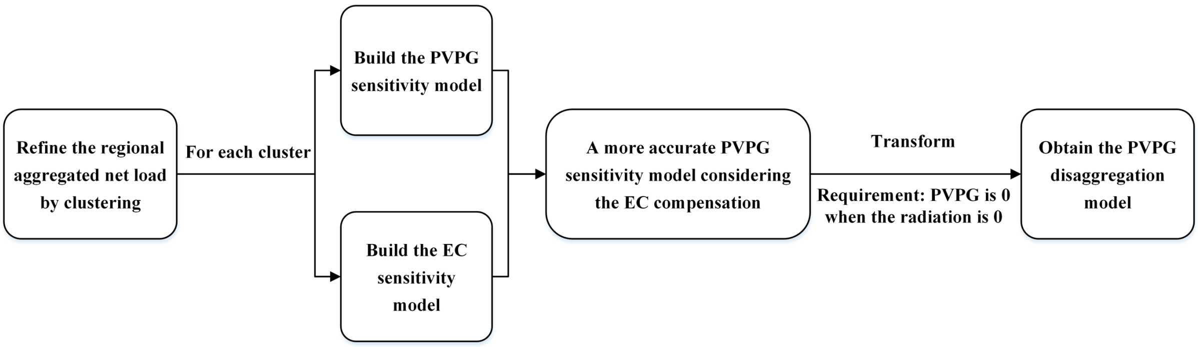

Refine the regional aggregated net load by clustering
For each cluster
Build the PVPG sensitivity model
Build the EC sensitivity model
A more accurate PVPG sensitivity model considering the EC compensation
Transform
Requirement: PVPG is 0 when the radiation is 0
Obtain the PVPG disaggregation model

Fig. 3 The framework of the proposed BtM net load disaggregation methodology

3.2. Net load disaggregation

The most difficult part of net load disaggregation is how to disaggregate PVPG sequences from net load without historical separated sub-meter data as reference. In this section, considering the relationships between the PVPG and solar irradiation, the PVPG sensitivity model is built by employing *k*-nearest neighbor (KNN) [23] and LSTM [24]. Furthermore, the EC model is established fully taking the strong correlation between ambient temperature and day type by using KNN and LSTM, too. The net load disaggregation model is proposed with the output characteristics of PV panels. The details are given as follows:

3.2.1. PVPG sensitivity model

Because BtM sub-meter data does not easily access from utility grid side, it is not feasible to obtain PVPG through supervised or semi-supervised methods. However, the approximate value of PVPG difference can be obtained by making a difference of a paired net load, which has similar EC behavior. For the sake of simplicity, net load difference ΔP_{N}, EC difference ΔP_{C} and PVPG difference ΔP_{G} between day i and day j are given by Eq. (4):

$$\begin{cases} \Delta P_{\mathrm{N}}^{i,j}\left(\mathbf{t}\right) = P_{\mathrm{N}}^{i}\left(\mathbf{t}\right) - P_{\mathrm{N}}^{j}\left(\mathbf{t}\right) \\ \Delta P_{\mathrm{C}}^{i,j}\left(\mathbf{t}\right) = P_{\mathrm{C}}^{i}\left(\mathbf{t}\right) - P_{\mathrm{C}}^{j}\left(\mathbf{t}\right) \\ \Delta P_{\mathrm{G}}^{i,j}\left(\mathbf{t}\right) = P_{\mathrm{G}}^{i}\left(\mathbf{t}\right) - P_{\mathrm{G}}^{j}\left(\mathbf{t}\right) \end{cases} \tag{4}$$

The estimated value of PVPG difference $\Delta\tilde{P}_{\mathrm{G}}$ between day i and day j is given by Eq. (5) below:

$$\Delta\tilde{P}_{\mathrm{G}}^{i,j}\left(\mathbf{t}\right) \approx -\Delta P_{\mathrm{N}}^{i,j}\left(\mathbf{t}\right) \tag{5}$$

$$s.t.\ \Delta P_{\mathrm{C}}^{i,j}\left(\mathbf{t}\right) \approx 0,\ \ \Delta P_{\mathrm{N}}^{i,j}\left(\mathbf{t}\right) = \Delta P_{\mathrm{C}}^{i,j}\left(\mathbf{t}\right) - \Delta P_{\mathrm{G}}^{i,j}\left(\mathbf{t}\right) \tag{6}$$

The load demand is determined by the customer behavior. Therefore, to search the dates with similar EC behavior, KNN is applied to find the most similar samples. In consideration of the PVPG concealing of the net load at time $\mathbf{t}_{\mathrm{day}}$, only the net load at time $\mathbf{t}_{\mathrm{night}}$ is chosen to find the nearest neighbors samples to avoid the distubance with the assumption that the days when customers with approximate nighttime EC behaviors are more likely to have approximate daytime EC behaviors. The matched EC day set of day i can be shown as below:

$$\mathbf{D}_{\mathrm{C}}^{i,k1} = \left\{\left(i,i_1\right),\left(i,i_2\right),\cdots,\left(i,i_l\right),\cdots,\left(i,i_k\right)\right\} \tag{7}$$

where $k1$ represents the number of nearest neighbors of PVPG sensitivity model.

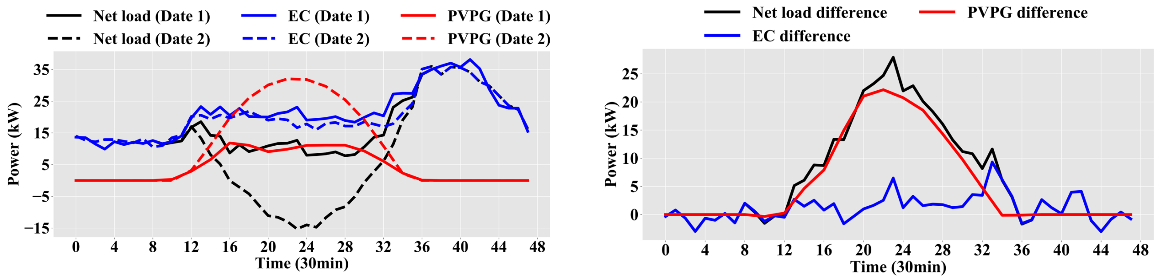
The set of estimated PVPG difference can be given by Eq. (8):

$$\Delta\tilde{\mathbf{P}}_{\mathrm{G}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right) \approx -\Delta\mathbf{P}_{\mathrm{N}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right) \tag{8}$$

$$s.t.\ \Delta P_{\mathrm{C}}^{i,i_l}\left(\mathbf{t}_{\mathrm{day}}\right) \approx 0,\ \ \Delta\mathbf{P}_{\mathrm{N}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right) = \Delta\mathbf{P}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right) - \Delta\mathbf{P}_{\mathrm{G}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right) \tag{9}$$

where $\Delta\mathbf{P}_{\mathrm{N}}$, $\Delta\mathbf{P}_{\mathrm{C}}$ and $\Delta\mathbf{P}_{\mathrm{G}}$ are the difference set of ΔP_{N}, ΔP_{C} and ΔP_{G}, respectively.

An example of obtained approximate PVPG difference is given in Fig. 4.



(a) Net load, EC and PVPG of a paring date with similar EC behaviors

(b) Net load, EC and PVPG difference of a paring date with similar EC behaviors

Fig. 4 An example of obtained approximate PVPG difference

Fig. 4(a) shows that by searching the date of similar EC behavior at time $\mathbf{t}_{\text{night}}$, two similar EC curves of a whole day $\mathbf{t}$ can be acquired. By making the difference between the net load of the matched two days, an approximate PVPG difference curve can be obtained as shown by the black solid line in Fig. 4(b).

By referring to the date index of Eq. (7), the corresponding difference set of DHI $\Delta\mathbf{I}_{\text{DHI}}^{\mathbf{D}_{\text{C}}^{i,k1}}$ and DNI $\Delta\mathbf{I}_{\text{DNI}}^{\mathbf{D}_{C}^{i,k1}}$ can be obtained by Eqs. (10) and (11):

$$\Delta\mathbf{I}_{\text{DHI}}^{\mathbf{D}_{\text{C}}^{i,k1}}\left(\mathbf{t}_{\text{day}}\right)=\left\{\Delta I_{\text{DHI}}^{i,i_1}\left(\mathbf{t}_{\text{day}}\right),\Delta I_{\text{DHI}}^{i,i_2}\left(\mathbf{t}_{\text{day}}\right),\cdots,\Delta I_{\text{DHI}}^{i,i_{k1}}\left(\mathbf{t}_{\text{day}}\right)\right\} \tag{10}$$

$$\Delta\mathbf{I}_{\text{DNI}}^{\mathbf{D}_{\text{C}}^{i,k1}}\left(\mathbf{t}_{\text{day}}\right)=\left\{\Delta I_{\text{DNI}}^{i,i_1}\left(\mathbf{t}_{\text{day}}\right),\Delta I_{\text{DNI}}^{i,i_2}\left(\mathbf{t}_{\text{day}}\right),\cdots,\Delta I_{\text{DNI}}^{i,i_{k1}}\left(\mathbf{t}_{\text{day}}\right)\right\} \tag{11}$$

where $\Delta I_{\text{GHI}}^{i,i_{k1}}$ is the GHI difference between day I_{GHI}^{i} and day $I_{\text{GHI}}^{i_{k1}}$, $\Delta I_{\text{GNI}}^{i,i_{k1}}$ is the GNI difference between day I_{GNI}^{i} and day $I_{\text{GNI}}^{i_{k1}}$.

As PVPG is mainly affected by DHI and DNI; LSTM, which has a good performance to deal with timing sequences problems, is applied to explore the nonlinear relationship between PVPG difference, DHI difference and DNI difference. The relationship between the above variables can be expressed by Eq. (12):

$$\Delta\tilde{\mathbf{P}}_{\text{G}}^{\mathbf{D}_{\text{C}}^{i,k1}}\left(\mathbf{t}_{\text{day}}\right)\approx-\Delta\mathbf{P}_{\text{N}}^{\mathbf{D}_{\text{C}}^{i,k1}}\left(\mathbf{t}_{\text{day}}\right)=f_{\text{I,G}}\left(\Delta\mathbf{I}_{\text{DNI}}^{\mathbf{D}_{\text{C}}^{i,k1}}\left(\mathbf{t}_{\text{day}}\right),\Delta\mathbf{I}_{\text{DHI}}^{\mathbf{D}_{\text{C}}^{i,k1}}\left(\mathbf{t}_{\text{day}}\right)\right) \tag{12}$$

where $f_{\text{I,G}}\left(\bullet\right)$ is the fitting function formed by LSTM between estimated PVPG and irradiation. For the sake of ensuring the number of features when training the LSTM model and reducing unnecessary information loss caused by the difference between the variables, the actual modeling is carried out in the form of Eq. (13):

$$\Delta\tilde{\mathbf{P}}_{\mathrm{G}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right)\approx-\Delta\mathbf{P}_{\mathrm{N}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right)=f_{\mathrm{I,G}}\left(\mathbf{I}_{\mathrm{DNI}}^{i}\left(\mathbf{t}_{\mathrm{day}}\right),\mathbf{I}_{\mathrm{DNI}}^{k1}\left(\mathbf{t}_{\mathrm{day}}\right),\mathbf{I}_{\mathrm{DHI}}^{i}\left(\mathbf{t}_{\mathrm{day}}\right),\mathbf{I}_{\mathrm{DHI}}^{k1}\left(\mathbf{t}_{\mathrm{day}}\right)\right) \tag{13}$$

$$s.t.\ \begin{cases}\mathbf{I}_{\mathrm{DNI}}^{i}\left(\mathbf{t}_{\mathrm{day}}\right)-\mathbf{I}_{\mathrm{DNI}}^{k1}\left(\mathbf{t}_{\mathrm{day}}\right)=\Delta\mathbf{I}_{\mathrm{DNI}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right)\\ \mathbf{I}_{\mathrm{DHI}}^{i}\left(\mathbf{t}_{\mathrm{day}}\right)-\mathbf{I}_{\mathrm{DHI}}^{k1}\left(\mathbf{t}_{\mathrm{day}}\right)=\Delta\mathbf{I}_{\mathrm{DNI}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right)\end{cases} \tag{14}$$

3.2.2. PVPG disaggregation model

With the help of the physical characteristics of PV panels that PVPG equals to 0 when DNI and DHI are 0, the disaggregated PVPG $P'^{i}_{\mathrm{PV}}(\mathbf{t}_{\mathrm{day}})$can be captured by Eq. (15) according to the relationship $f_{\mathrm{I,G}}\left(\bullet\right)$derived in Eq. (12) by only inputting DNI and DHI of day i:

$$P'^{i}_{\mathrm{G}}(\mathbf{t}_{\mathrm{day}})-0=f_{\mathrm{I,G}}((I_{\mathrm{DNI}}^{i}(\mathbf{t}_{\mathrm{day}})-0),(I_{\mathrm{DHI}}^{i}(\mathbf{t}_{\mathrm{day}})-0)) \tag{15}$$

Except for DNI and DHI, the operating temperature of PV panels will also dramatic influence PV conversion efficiency [25]. Therefore, ambient temperature should be added when building the PVPG sensitivity model in Section 3.2.1. It is important to note that, unlike the relationship between PVPG and irradiation, we cannot obtain the exact value of EC when the temperature is a specific value. In other words, when the PVPG sensitivity model is transformed into PVPG disaggregation model, only the ambient temperature of the required PVPG disaggregation day can be used. Hence, Eq. (12) should be rewritten as Eq. (16) instead of Eq. (17).

$$\Delta\tilde{\mathbf{P}}_{\mathrm{G}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right)\approx-\Delta\mathbf{P}_{\mathrm{N}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}=f_{\mathrm{I,G}}\left(\Delta\mathbf{I}_{\mathrm{DHI}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right),\Delta\mathbf{I}_{\mathrm{DNI}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right),\mathbf{T}^{i}\left(\mathbf{t}_{\mathrm{day}}\right)\right) \tag{16}$$

$$\Delta\tilde{\mathbf{P}}_{\mathrm{G}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right)\approx-\Delta\mathbf{P}_{\mathrm{N}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}=f_{\mathrm{I,G}}\left(\Delta\mathbf{I}_{\mathrm{DNI}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right),\Delta\mathbf{I}_{\mathrm{DHI}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right),\Delta\mathbf{T}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right)\right) \tag{17}$$

where $\mathbf{T}$ is the set of ambient temperature T. The PVPG disaggregation model transformed from the PVPG sensitivity model is written as Eq. (18) below:

$$\mathbf{P}'^{i}_{\mathrm{G}}(\mathbf{t}_{\mathrm{day}})-\mathbf{0}=f_{\mathrm{I,G}}\left(\mathbf{I}_{\mathrm{DNI}}^{i}\left(\mathbf{t}_{\mathrm{day}}\right),\mathbf{0},\mathbf{I}_{\mathrm{DHI}}^{i}\left(\mathbf{t}_{\mathrm{day}}\right),\mathbf{0},\mathbf{T}^{i}\left(\mathbf{t}_{\mathrm{day}}\right)\right) \tag{18}$$

3.2.3. EC sensitivity model

The disaggregation model built with Eq. (18) based on the approximate estimated PVPG difference according to Eq. (8) is still flawed, because it is almost impossible to find these matched days that have completely consistent EC curves, especially for residential load in which the EC behavior is changeable. Therefore, the values of Eq. (8) include the EC difference as shown in Fig. 4(b) with the blue solid line between the matched days, and its actual representation should be given by Eq. (19):

$$\Delta\tilde{\mathbf{P}}_{\mathrm{G}}^{\mathbf{D}_{\mathrm{C}}^{i,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)=-\Delta\mathbf{P}_{\mathrm{N}}^{\mathbf{D}_{\mathrm{C}}^{i,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)+\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{C}}^{i,k2}}\left(\mathbf{t}_{\mathrm{day}}\right) \tag{19}$$

where $k2$ represents the number of nearest neighbors of EC sensitivity model, $\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{C}}^{i,k2}}$ is the compensation EC difference of matched EC day set $\mathbf{D}_{\mathrm{C}}^{i,k2}$.

To build a more accurate disaggregation model, the values of compensation EC difference $\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{C}}^{i,k2}}$ must be acquired. Similar to the PVPG data, in the case of BtM, EC data of sub-meter is inaccessible directly. However, the estimated EC difference can be obtained by making a difference of a paired net load, which has similar PVPG. Considering that for the customers, there is generally no physical configuration change to the PV devices in a short period of time, by searching the days with similar DNI and DHI for the same customers, the paired day set with similar PVPG $\mathbf{D}_{\mathrm{G}}^{j,k2}$ can be obtained using KNN.

$$\mathbf{D}_{\mathrm{G}}^{j,k2}=\left\{\left(j,j_{1}\right),\left(j,j_{2}\right),\cdots,\left(j,j_{l}\right),\cdots,\left(j,j_{k2}\right)\right\} \tag{20}$$

The set of estimated EC difference $\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}$ can be written as Eq. (21):

$$\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)\approx\Delta\mathbf{P}_{\mathrm{N}}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}\left(\mathbf{t}_{\mathrm{day}}\right) \tag{21}$$

$$s.t.\ \Delta P_{\mathrm{G}}^{j,j_{l}}\left(\mathbf{t}_{\mathrm{day}}\right)\approx 0,\ \ \Delta\mathbf{P}_{\mathrm{N}}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)=\Delta\mathbf{P}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)-\Delta\mathbf{P}_{\mathrm{G}}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}\left(\mathbf{t}_{\mathrm{day}}\right) \tag{22}$$

According to [26], the change of EC for residents is mainly driven by the ambient temperature. Hence, the EC sensitivity model can be built by $\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}$ and ambient temperature difference $\Delta\mathbf{T}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}$ of matched PVPG day set $\mathbf{D}_{\mathrm{G}}^{j,k2}$ from Eq. (23):

$$\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)=f_{\mathrm{T,C}}\left(\Delta\mathbf{T}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)\right) \tag{23}$$

$$\Delta\mathbf{T}^{\mathbf{D}_{\mathrm{G}}^{j,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)=\left\{\Delta T^{j,j_{1}}\left(\mathbf{t}_{\mathrm{day}}\right),\Delta T^{j,j_{2}}\left(\mathbf{t}_{\mathrm{day}}\right),\cdots,\Delta T^{j,j_{k2}}\left(\mathbf{t}_{\mathrm{day}}\right)\right\} \tag{24}$$

where $f_{\mathrm{T,C}}\left(\bullet\right)$ is the fitting function formed by LSTM between estimated EC and ambient temperature.

The compensation EC difference $\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{C}}^{i,k2}}$ can be calculated with corresponding temperature difference $\Delta\mathbf{T}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{C}}^{i,k2}}$ from the EC sensitivity model as shown in Eq. (25):

$$\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{C}}^{i,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)=f_{\mathrm{T,C}}\left(\Delta\mathbf{T}^{\mathbf{D}_{\mathrm{C}}^{i,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)\right) \tag{25}$$

Based on the same reason by rewriting Eq. (12) to Eq. (13), the EC sensitivity modeling is carried out in the form of Eq. (26):

$$\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{C}}^{i,k2}}\left(\mathbf{t}_{\mathrm{day}}\right)=f_{\mathrm{T,C}}\left(\mathbf{T}^{i}\left(\mathbf{t}_{\mathrm{day}}\right),\mathbf{T}^{k2}\left(\mathbf{t}_{\mathrm{day}}\right)\right) \tag{26}$$

It is important to note the monotonicity of EC sensitivity model and PVPG sensitivity model. PVPG and solar irradiation are always positively correlated, while EC and ambient temperature are positively correlated in summer, negatively correlated in winter. Consequently, when matching the days with similar DNI and DHI, the search range should be within the same monotonicity of the EC relative ambient temperature. Considering that the shape of EC curve varies greatly in different seasons, before building the EC sensitivity model, the net load data is clustered to ensure the search range.

Since the daytime net load is composed of EC and PVPG, it is difficult to ensure the clustering results only reflect the differences in EC behavior and not affected by the variability of PVPG. In order to solve the above problems, evening net load containing only pure EC at $\mathbf{t}_{\mathrm{night}}$ is used for clustering based on the continuity of customers' EC behavior. By assuming that if the evening EC has an approximate distribution, the daytime load also has an approximate distribution. Three clustering methods including Gaussian mixture models (GMM) [27], Fuzzy C-means (FCM) with Euclidean distance (ED) and FCM with dynamic time warping (DTW) [28], [29] are employed for experiments and comparisons in case study.

3.2.4. PVPG disaggregation model with EC difference compensation

After the compensation EC difference $\Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}$ is obtained through the EC sensitivity model, a more precise PVPG sensitivity model can be rewritten from Eq. (16) as:

$$\Delta\tilde{\mathbf{P}}_{\mathrm{G}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right) \approx \Delta\tilde{\mathbf{P}}_{\mathrm{C}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}} - \Delta\mathbf{P}_{\mathrm{N}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}} = f_{\mathrm{I,G}}^{'}\left(\Delta\mathbf{I}_{\mathrm{DNI}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right), \Delta\mathbf{I}_{\mathrm{DHI}}^{\mathbf{D}_{\mathrm{C}}^{i,k1}}\left(\mathbf{t}_{\mathrm{day}}\right), \mathbf{T}^{i}\left(\mathbf{t}_{\mathrm{day}}\right)\right) \tag{27}$$

where $f_{\mathrm{I,G}}^{'}(\bullet)$ is the compensatory fitting function formed by LSTM between estimated PVPG and irradiation.

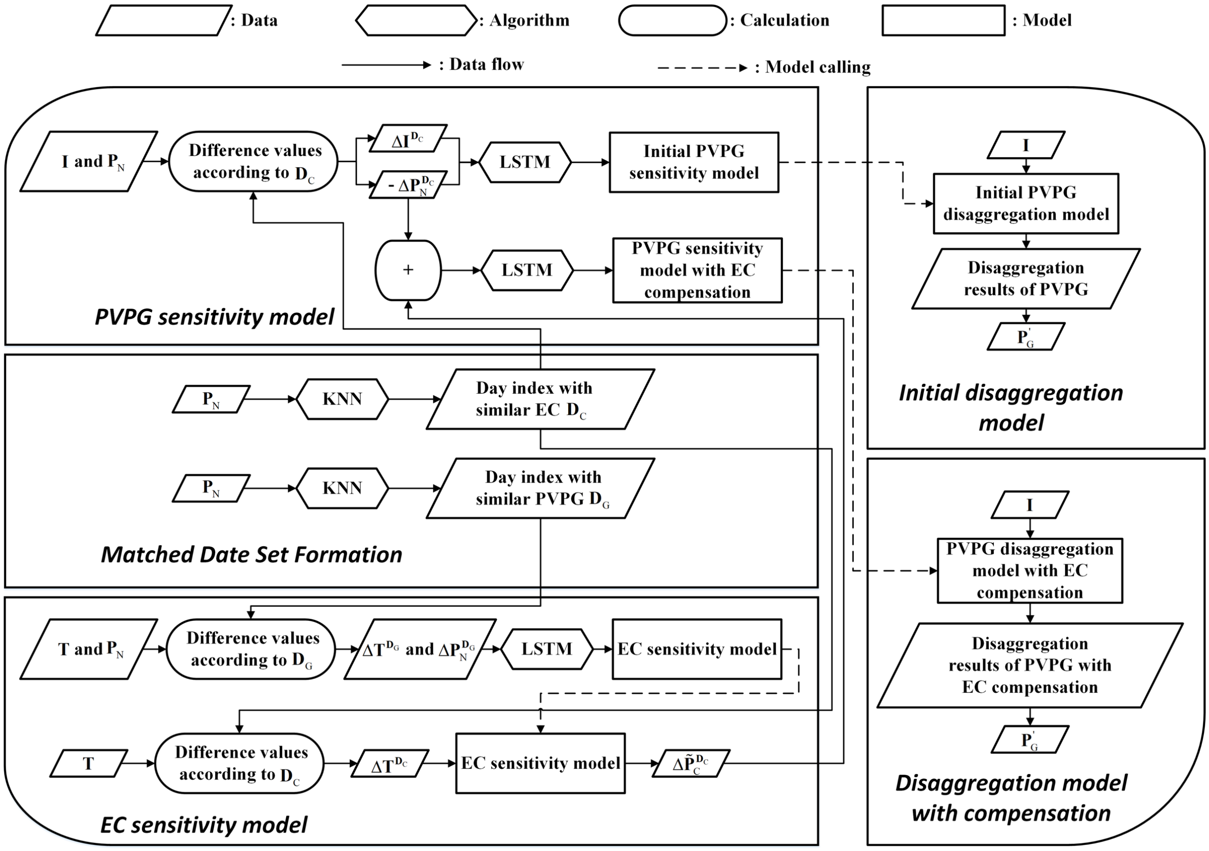
The final PVPG disaggregation model in practice can be converted from compensated PVPG sensitivity model to Eq. (28):

$$\mathbf{P}_{\mathrm{G}}^{'i}(\mathbf{t}_{\mathrm{day}}) - \mathbf{0} = f_{\mathrm{I,G}}^{'}(\mathbf{I}_{\mathrm{DNI}}^{i}(\mathbf{t}_{\mathrm{day}}), \mathbf{0}, \mathbf{I}_{\mathrm{DHI}}^{i}(\mathbf{t}_{\mathrm{day}}), \mathbf{0}, \mathbf{T}^{i}(\mathbf{t}_{\mathrm{day}})) \tag{28}$$

The disaggregated EC $P_{\mathrm{C}}^{'i}(\mathbf{t}_{\mathrm{day}})$ in day i is further disaggregated as below:

$$P_{\mathrm{C}}^{'i}(\mathbf{t}_{\mathrm{day}}) = P_{\mathrm{N}}^{i}(\mathbf{t}_{\mathrm{day}}) + P_{\mathrm{G}}^{'i}(\mathbf{t}_{\mathrm{day}}) \tag{29}$$

The model of the proposed net load disaggregation method is illustrated in Fig. 5:



: Data
: Algorithm
: Calculation
: Model
: Data flow
: Model calling
I and P_N
Difference values according to D_C
ΔI^{D_C}
- ΔP_N^{D_C}
LSTM
Initial PVPG sensitivity model
+
LSTM
PVPG sensitivity model with EC compensation
PVPG sensitivity model
P_N
KNN
Day index with similar EC D_C
P_N
KNN
Day index with similar PVPG D_G
Matched Date Set Formation
T and P_N
Difference values according to D_G
ΔT^{D_G} and ΔP_N^{D_G}
LSTM
EC sensitivity model
T
Difference values according to D_C
ΔT^{D_C}
EC sensitivity model
ΔP̃_C^{D_C}
EC sensitivity model
I
Initial PVPG disaggregation model
Disaggregation results of PVPG
P'_G
Initial disaggregation model
I
PVPG disaggregation model with EC compensation
Disaggregation results of PVPG with EC compensation
P'_G
Disaggregation model with compensation

Fig. 5. The proposed disaggregation model

As shown in Fig. 5, the proposed BtM disaggregation model does not involve any PV system geometry and PV parameter assumptions. This means that it is completely data-driven. The distinction of the feature requirements in different stages is highlighted in Table 3.

Table 3
Method and features requirements of different stages.

Stages	Features				
	Net Load	Ambient Temperature	One-sided Ambient Temperature	DNI	DHI
PVPG Sensitivity Modelling	√	×	√	√	√
EC Sensitivity Modelling	√	√	×	×	×

4. Performance metrics and method comparisons

4.1. Clustering assessment metrics

Before building the EC sensitivity model, it is essential to cluster the net load. Davies-Bouldin index (DBI) is employed to reflect the quality of the clustering by measuring the mean of the maximum similarity of each cluster [30]. It should be noted that when using DBI for FCM-DTW, the intra-class distance and the inter-class distance should be calculated according to DTW distance. The clustering performance is better with a smaller DBI value.

4.2. Net load disaggregation assessment metrics

The accuracy of the proposed net load disaggregation method is evaluated by root mean square error (RMSE) and the designed mean absolute percentage error (MAPE).

Hourly RMSE is defined in Eq. (30) for evaluating the disaggregated PVPG only, because the disaggregated EC is obtained by the difference between the net load and the disaggregated generation.

$$RMSE_{\mathrm{G}}^{i}=\sqrt{\frac{1}{t_{\mathrm{e}}-t_{\mathrm{s}}}\sum_{t\in\mathbf{t}_{\mathrm{day}}}\left(P_{\mathrm{G}}^{\prime i}\left(t\right)-P_{\mathrm{G}}^{i}\left(t\right)\right)^{2}} \tag{30}$$

As indicated in [11] and [13], hourly MAPE is highly sensitive to the time interval of low absolute PVPG. To avoid the above problems, by referring to [11], we also use only the middle part of daytime (from 10:00 to 15:00) to calculate the hourly MAPE. The hourly MAPE of estimating PVPG is given by Eq. (31).

$$MAPE_{\mathrm{G}}^{i}=\frac{100\%}{\Delta t_{\mathrm{middle}}}\sum_{t\in\mathbf{t}_{\mathrm{middle}}}\left|\frac{P_{\mathrm{G}}^{\prime i}\left(t\right)-P_{\mathrm{G}}^{i}\left(t\right)}{P_{\mathrm{G}}^{i}\left(t\right)}\right| \tag{31}$$

where $\mathbf{t}_{\mathrm{middle}}$ is the set of the middle part of $\mathbf{t}_{\mathrm{day}}$, $\Delta t_{\mathrm{middle}}$ is the duration time of $\mathbf{t}_{\mathrm{middle}}$.

4.3. Experiment comparisons

In order to analyze the impact of data partition on building the net load disaggregation model, GMM, FCM-ED and FCM-DTW are employed to make a comparison. The core of the disaggregation algorithm proposed in this paper is the construction of the sensitivity models. Taking into account the complex nonlinear relationship between variables, the fitting of the variables is completed by machine learning. The $k1$ and $k2$ values of KNN determine the size of the dataset for building PVPG sensitivity model and EC sensitivity model, hence the disaggregation results under different combination of k and $k2$ are discussed. In order to illustrate the general applicability of the sensitivity model and the superiority of LSTM for processing this problem, multilayer perceptron (MLP) [36] is introduced as a benchmark for comparison. Disaggregation with or without compensation are compared to demonstrate the effectiveness of the introduction of EC sensitivity model, and two unsupervised net load disaggregation methods from [14] and [15] are reproduced and used to compare with the proposed disaggregation method. In addition, the robustness of the proposed algorithm is illustrated by the disaggregation results obtained by building a model using meteorological data collected from heterogeneous sources and the disaggregation results obtained from heatwave condition.

5. Case study

5.1. Experimental platform and model parameter setting

All experiments were implemented by Python 3.7.7 on a server with NVIDIA Geforce RTX 2080Ti GPU and 64 GB of RAM.

Data from Ausgrid provides the separate timing series of PVPG and EC, so the separated data can be used for experimental verification, but not using in BtM PVPG disaggregation modelling. The data of 1096 days in the three years (including the 17 extreme weather days) described in Section 2.2 are all involved in the construction of the disaggregation model. Considering the daytime hours are different every day, in order to ensure uniformity for modelling, the earliest moment with PVPG occurrence at 5:00 and the latest moment with PVPG disappearance at 19:30 are set as the sunrise time t_s and the sunset time t_e, respectively.

Hyper-parameters of the artificial neural network are shown in Appendix B. Considering the length of this paper, the analysis of some experimental results is only carried out for Sub-region 2.

5.2. Data set partitions and net load matching

In order to exclude the measurement errors caused by the mismatch between observes meteorological and PVPG to better analyze the data set partitions and net load matching of the proposed method, experiments of this section are performed using semi-synthetic data.

5.2.1. Analysis of the data set partitions

When building the proposed disaggregation model, clustering is employed to partition net load data. Considering that the quantity and quality of proposed disaggregation model depend on the clustering results, three clustering methods including FCM based on ED, FCM based on DTW and GMM of clustering number c ranges from 2 to 4 are implemented and compared. In the data partition analysis, for variable control, the nearest neighbor values of PVPG sensitivity model is set to 10, the nearest neighbor values of EC sensitivity model is set to 30. Table 4 shows the DBI of different clustering methods.

Table 4
DBI of different clustering methods.

Method	**Number of clusters**		
	2	**3**	**4**
FCM-ED	0.638	0.933	1.219
FCM-DTW	0.588	0.858	1.131
GMM	0.797	1.352	1.620

All three methods achieve the best clustering results when the number of clusters is equal to 2. The DBI of FCM-DTW is calculated based on the DTW distance, so the value is the smallest. The disaggregation results using MLP and LSTM of different parameter combinations are presented in Tables 5 and 6.

Table 5

Average hourly MAPE of PVPG disaggregation with different clustering methods based on MLP for 3 years.

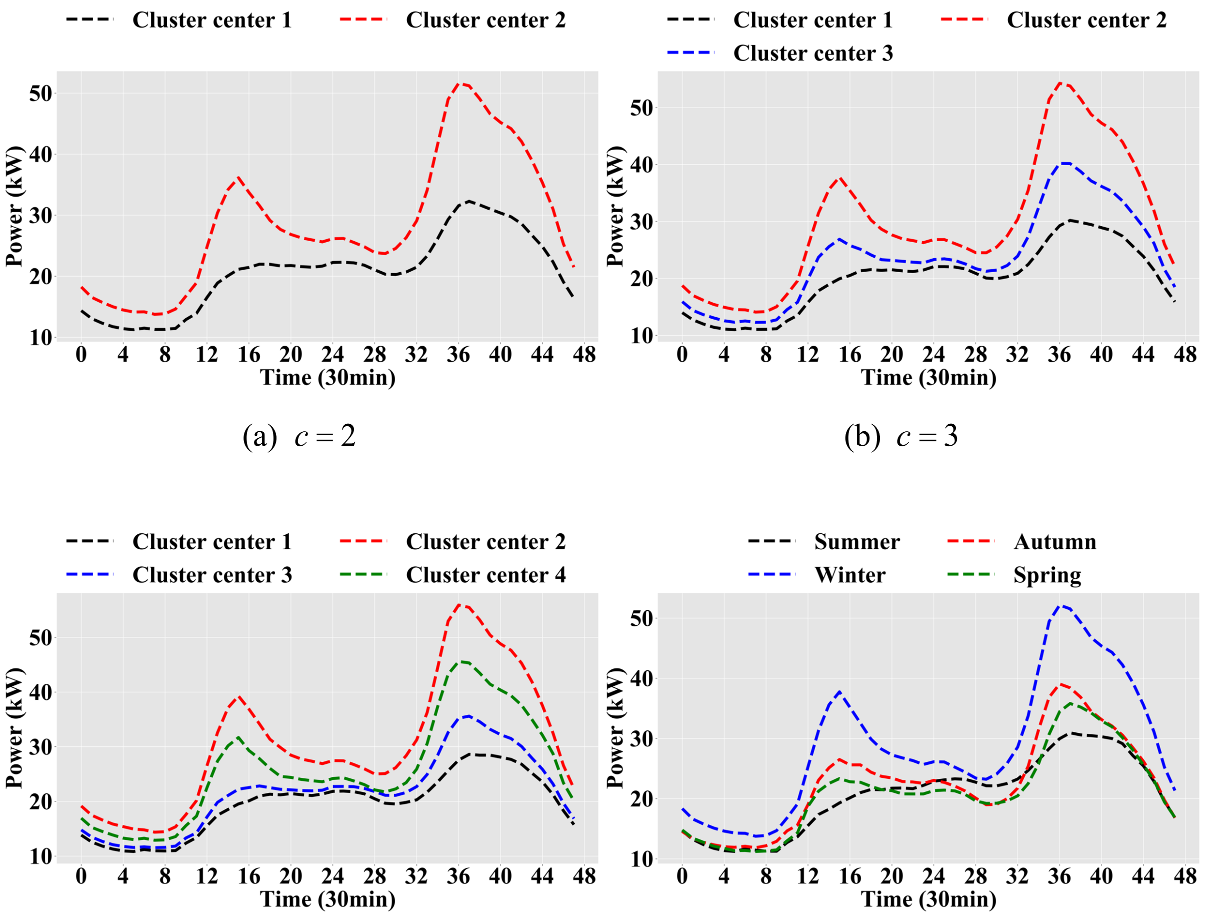
Number of clustering	Uncompensated (%)			Compensated (%)		
	FCM-ED	FCM-DTW	GMM	FCM-ED	FCM-DTW	GMM
2	13.03	11.83	13.71	**5.08**	5.18	6.50
3	14.90	13.43	11.54	8.00	7.26	6.67
4	15.49	15.41	14.43	8.16	7.52	8.67
No seasonal partition		14.66			8.23	
Seasonal partition		14.72			7.55	

Table 6

Average hourly MAPE of PVPG disaggregation with different clustering methods based on LSTM for 3 years.

Number of clustering	Uncompensated (%)			Compensated (%)		
	FCM-ED	FCM-DTW	GMM	FCM-ED	FCM-DTW	GMM
2	10.55	10.57	12.76	**3.20**	3.96	5.12
3	12.88	12.66	11.96	6.90	6.10	8.29
4	13.51	14.10	13.16	6.62	6.49	6.49
No seasonal partition		13.61			5.59	
Seasonal partition		13.56			5.82	

Tables 5 and 6 show that for each clustering method, the optimal disaggregation result appears when the number of clusters is 2. The observation is consistent with the result of the optimal number of clusters, while the disaggregation with no seasonal partition does not achieve a better accuracy. This can be interpreted as the influence of temperature on EC presents different correlations in different seasons, and the correlation is completely opposite in winter and summer. With clustering, the temperature-driven EC is divided into two types with obvious morphological differences, so that the sensitivity model can be more important. In contrast, the clustering method does not have a large impact on the decomposition accuracy, while the number of clusters has a greater impact. Therefore, FCM-ED is used to analyze the impact of the number of clusters on the proposed disaggregation model. The class center and the corresponding date distribution is shown in Figs. 6 and 7, respectively.



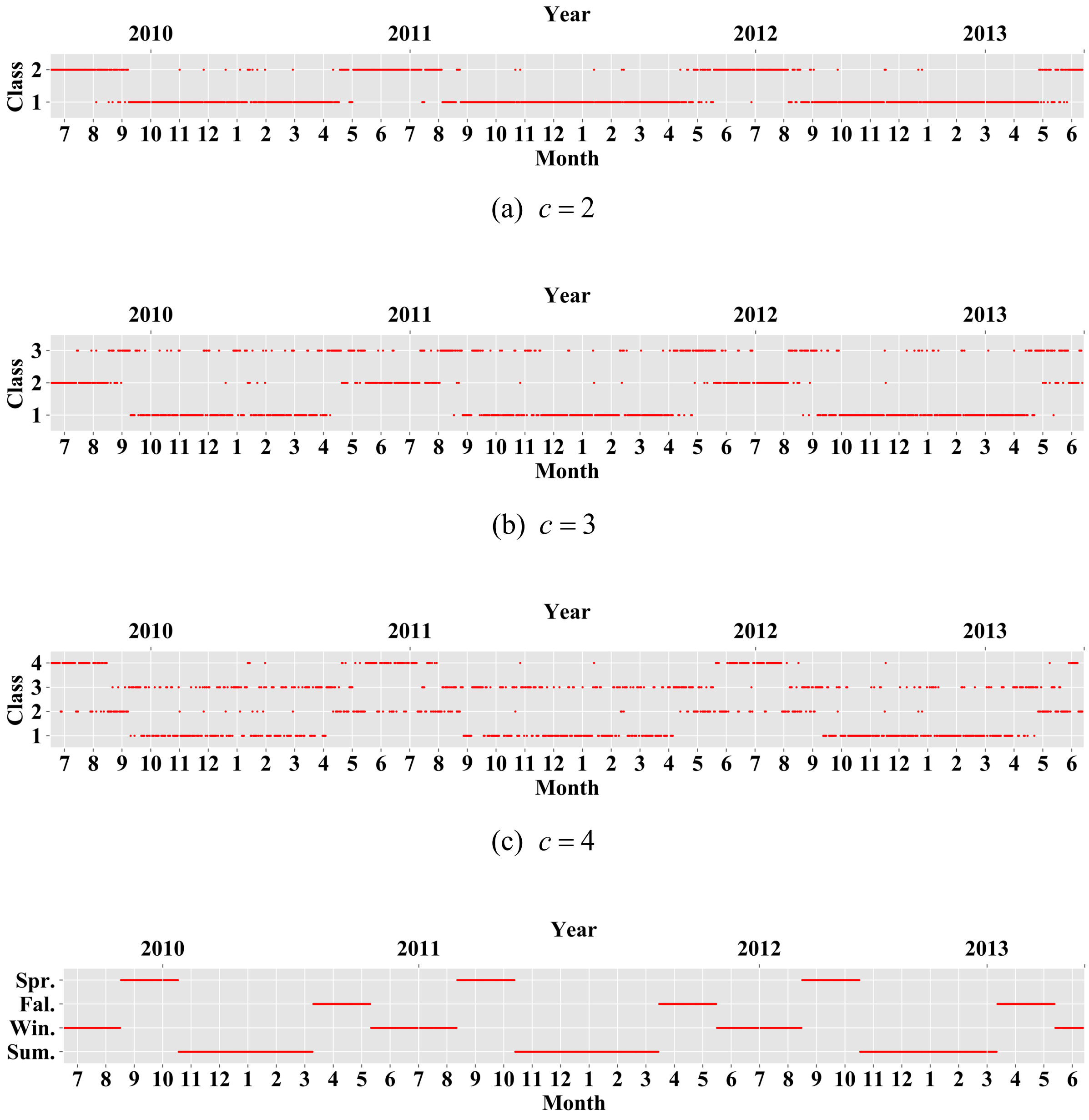
(a) $c = 2$

(b) $c = 3$

(c) $c = 4$

(d) Seasonal partition

Fig. 6 Class centers of the data partition



(a) $c = 2$

(b) $c = 3$

(c) $c = 4$

(d) Seasonal partition

Fig. 7 Date distribution of data partition

The cluster center in Fig. 6 is obtained from the clustering of EC curves at nighttime $\mathbf{t}_{\text{night}}$, but displayed as a whole day time $\mathbf{t}$ to facilitate analysis. From the results of clustering centers, it is feasible to pair or cluster through local nighttime curves to reflect the global EC behavior.

Average ambient temperature in summer in Sydney is usually not higher than 26°C, so the cooling load is less, which is represented as a black dotted line with a lower amplitude in Fig. 6(a). Considering the demand for heating is large in winter with average ambient temperature at about $13\,^{\circ}\mathrm{C}$, the load amplitude is relatively high, which is represented by the red dotted line in Fig. 6(a). However, Figs. 6(b) and 6(c) show that when the number of clusters is set to 3 or 4, the newly added cluster centers are between the typical winter and summer curves, and the centers characteristics are not obvious. The DBI index can also prove that the clustering quality is poor. Unreasonable data division will increase the number of EC sensitive models that need to be built, which may result in a decrease in the amount of model training data and affect model performance. Tables 5 and 6 also show the disaggregation of simple season division. The disaggregation result of seasonal division is similar to those with clusters number of 3, but it is better than the number of clusters of 4. From Fig. 6(d), it can be seen that in addition to the winter curve center, the difference between the center of the remaining season is chaotic for the season division, but the data distribution does not destroy the continuity of the time series shown in Fig. 7(d). In our experiments, it is found that the continuity of the data has a greater impact on the quality of the PVPG sensitivity model. This may be due to the input continuity of DNI and DHI, which can more regularly reflect the elevation and azimuth of the sun.

5.2.2. Analysis of the net load matching

In Section 3.2.1, when building the PVPG sensitivity model, the dates with similar EC behavior at time $\mathbf{t}_{\mathrm{day}}$ is determined by the net load at $\mathbf{t}_{\mathrm{night}}$ using KNN. Also, in Section 3.2.3, when building the EC sensitivity model, the dates with similar PVPG is determined by the matched DNI and DHI. To illustrate the effectiveness of the proposed strategy, the similarity corresponding to the kth match when finding the approximate EC behavior, PVPG by nighttime EC, and irradiation (DNI and DHI), is given in Fig. 8. The similarity is measured in terms of hourly MAPE and the results are shown as an average values of 1096 days for 3 years with the $k1$ and $k2$ range from 1-200.

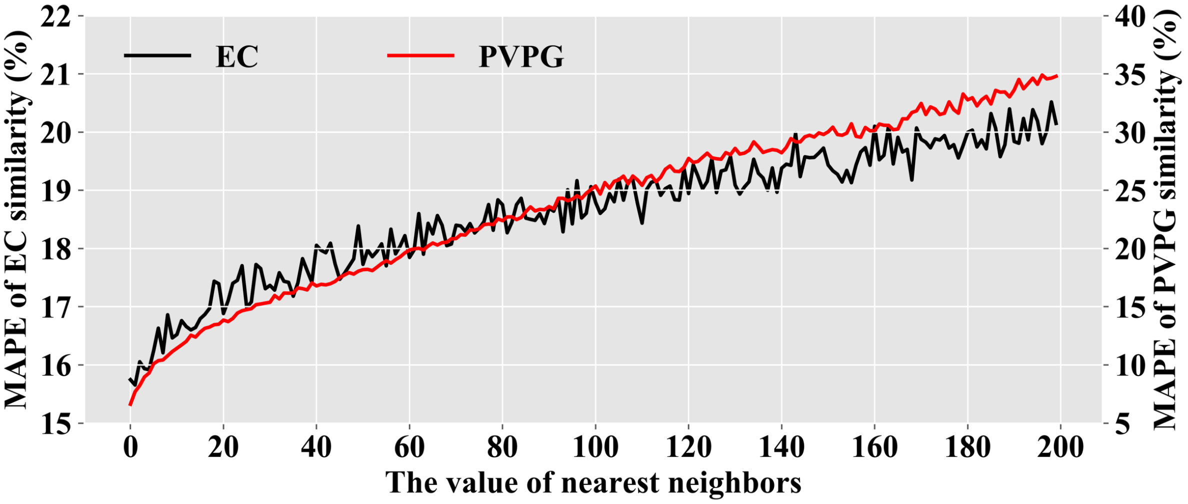
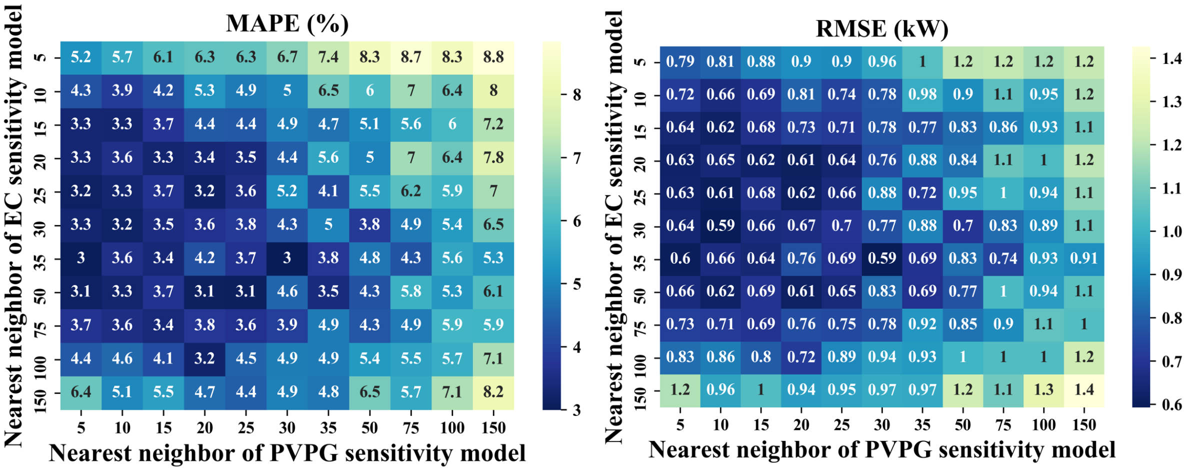


Fig. 8 The average similarity of matched EC and PVPG with different k

The results indicate that it is feasible to perform similarity matching with such strategies, and the variability of both the matched EC and PVPG shows an overall increasing trend with increasing k of nearest neighbors. The variability value of PVPG starts at a small amount, which means that for smaller values of $k2$, finding the net load with similar DNI and DHI by KNN and making a difference, can result in a sequence containing almost only EC differences information. In contrast, the variability value of EC starts at a large amount, which means that for smaller values of $k1$, finding the net load with similar nighttime EC behavior by KNN and making a difference, does not result in a sequence containing only PVPG differences information. This is the main source of the disaggregation error of the initial PVPG disaggregation model, and the compensation by building the EC sensitivity model is precisely to correct this part of the error.

Fig. 9 shows the effect of different nearest neighbor values on the disaggregation accuracy when FCM-ED is used with 2 as the number of clusters.



(a) Average hourly MAPE for 3 years

(b) Average hourly RMSE for 3 years

Fig. 9 The disaggregation error under different k combinations

In Fig. 9, when the combinations of $k1$ values of PVPG sensitivity model vary from 5 to 25 and $k2$ values of EC sensitivity model vary from 20 and 50, the proposed disaggregation method possesses comparatively small MAPE and RMSE. It can be seen that when $k2$ is a constant, both the average hourly MAPE and RMSE for 3 years in the heat map become larger as $k1$ increases. This is because when building the PVPG sensitivity model using KNN to find the dates with similar EC behaviors, the EC difference error introduced is already minimized. With the increase of $k1$, this error will gradually accumulate and lead to the decrease of disaggregation accuracy. In addition, the PVPG is directly affected by the irradiation, and the relationship between the two can be captured by the proposed machine learning model with a small amount of data, which is also the reason for the higher accuracy when the value of $k1$ is smaller. When $k1$ is a constant, both the average hourly MAPE and RMSE for 3 years in the heat map shows a trend from decline to rise as $k2$ increases. This may due to the relationship between EC behavior and temperature is complicated. Trying to reflect the EC difference through the temperature difference as effectively as possible requires a larger amount of data for machine learning modelling, which is the reason for the error decreases gradually at the beginning as $k2$ increases. However, excessively increase in $k2$ will also cause the interference information of PVPG difference accumulates, leading to the EC sensitivity model cannot correctly reflect the relationship between EC and temperature. This is the reason why the error rises again.

Overall, $k1$ needs to be kept at a small value, while $k2$ is a value that needs to be traded off to minimize the interference information of PVPG difference introduced while ensuring to have a sufficient amount of data to build the EC sensitivity model. It is notable from Fig. 9 that more satisfactory disaggregation results are obtained even for non-optimal combinations of $k1$ and $k2$. The subsequent experimental results in this paper are obtained based on $k1$ of 10 and $k2$ of 30.

5.3. Comparisons with state-of-the-art disaggregation methods

In this section, the performance of two unsupervised disaggregation methods reproduced from [14] and [15] are compared with the proposed method. The box plots of average hourly MAPE comparisons for 3 years are shown in Fig. 10. All the proposed methods are with EC difference compensation.

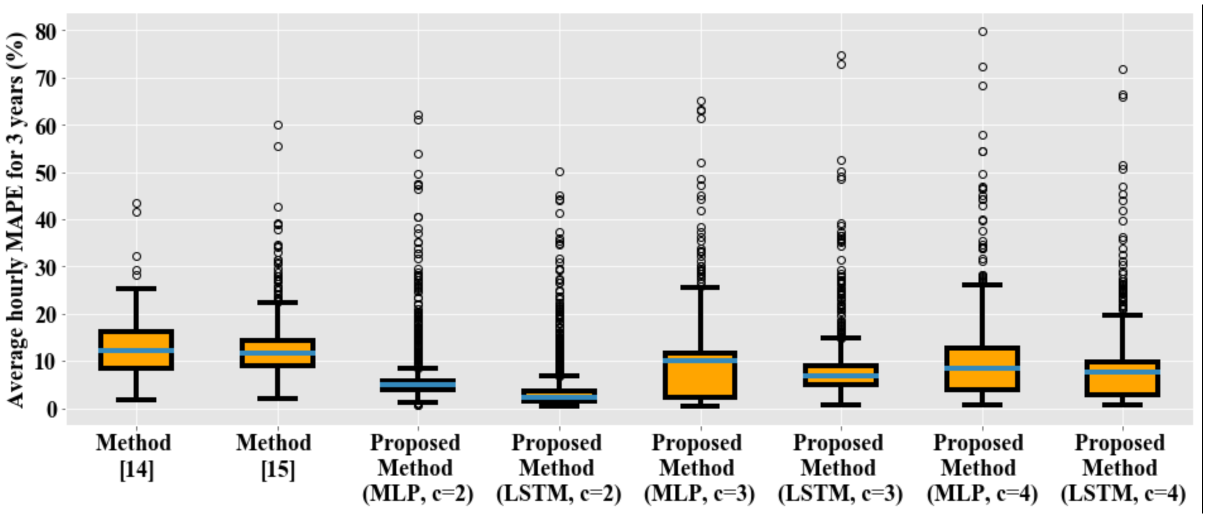


Fig. 10 Average hourly disaggregation MAPE comparison for 3 years

With this experimental data set, method in [14] has the worst disaggregation performance, the proposed method with compensation has better disaggregation results under different parameter combinations compared to it. Method in [15] has slightly higher median and box width with the proposed method with $c = 3$ or $c = 4$, but has less outliers. When $c = 2$, both the proposed methods constructed by MLP and LSTM have significant disaggregation accuracy improvements compared to methods from [14] and [15]. The use of MLP modeling can also achieve better disaggregation results as compared to the state-of-the-art methods, which can explain the general applicability of using neural networks to the disaggregation architecture to some extent. Fig. 10 shows that under the same number of clusters, the disaggregation effect of the model built by LSTM is better than that of MLP, which is mainly due to the advantages of LSTM in dealing with time series problems. Therefore, if a neural network with stronger learning ability can be used, the disaggregation accuracy may be further improved, which also shows the scalability of the proposed BtM PVPG disaggregation architecture.

Hourly MAPE profile for each day of the whole year 2012 with different methods is shown in Fig. 11.

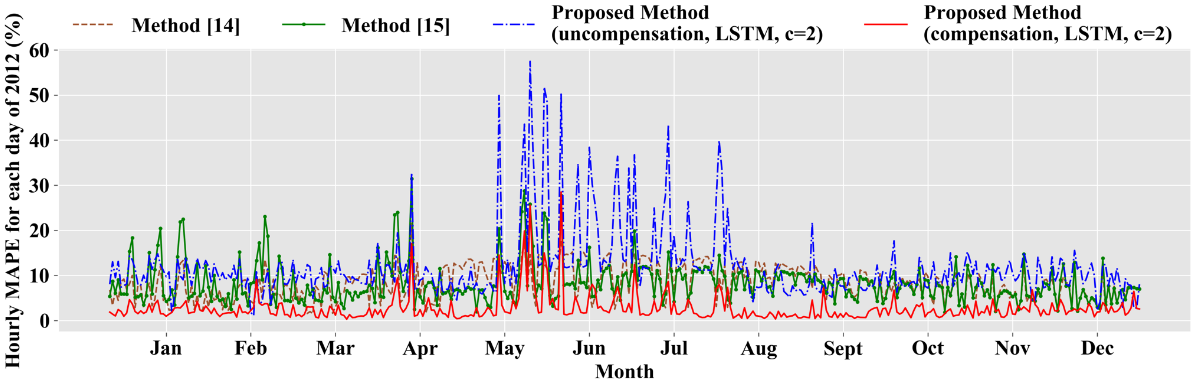


Fig. 11 Hourly MAPE profile for each day of year 2012 with different methods

In Fig. 11, proposed method constructed by LSTM, with and without compensation, $c = 2$ are compared. Method in [14] is designed based on a large number of PV system geometry assumptions. Therefore, if there is a discrepancy between the assumption and the actual situation, for example, the degradation of the PV arrays, the disaggregation error cannot be eliminated, and this kind of error exists at all times of the year. Method in [15] transforms the disaggregation problem into a constrained optimization problem with relevant constraints. However, in this method, the constraint conditions do not reflect the influence of temperature on the EC behavior, so there will be a large disaggregation error. The proposed method without compensation does not consider the changes in EC behavior caused by temperature when modelling, causing the net load difference in this case to be erroneously attributed to the difference in irradiation, so it has worse performance in winter when the temperature-sensitive load is more.

In contrast, the proposed method with compensation maintains a lower error range throughout the year, even in months when the temperature-sensitive load has a greater variation. Tables 5 and 6 also conclude that no matter what the combination of model parameters is, the compensated disaggregation model will have a huge improvement in accuracy as compared to the uncompensated disaggregation model.

The average hourly RMSE and MAPE of different disaggregation methods for 3 years are shown in Table 7.

Table 7
The average hourly RMSE and MAPE reduction for 3 years as compared to other algorithms.

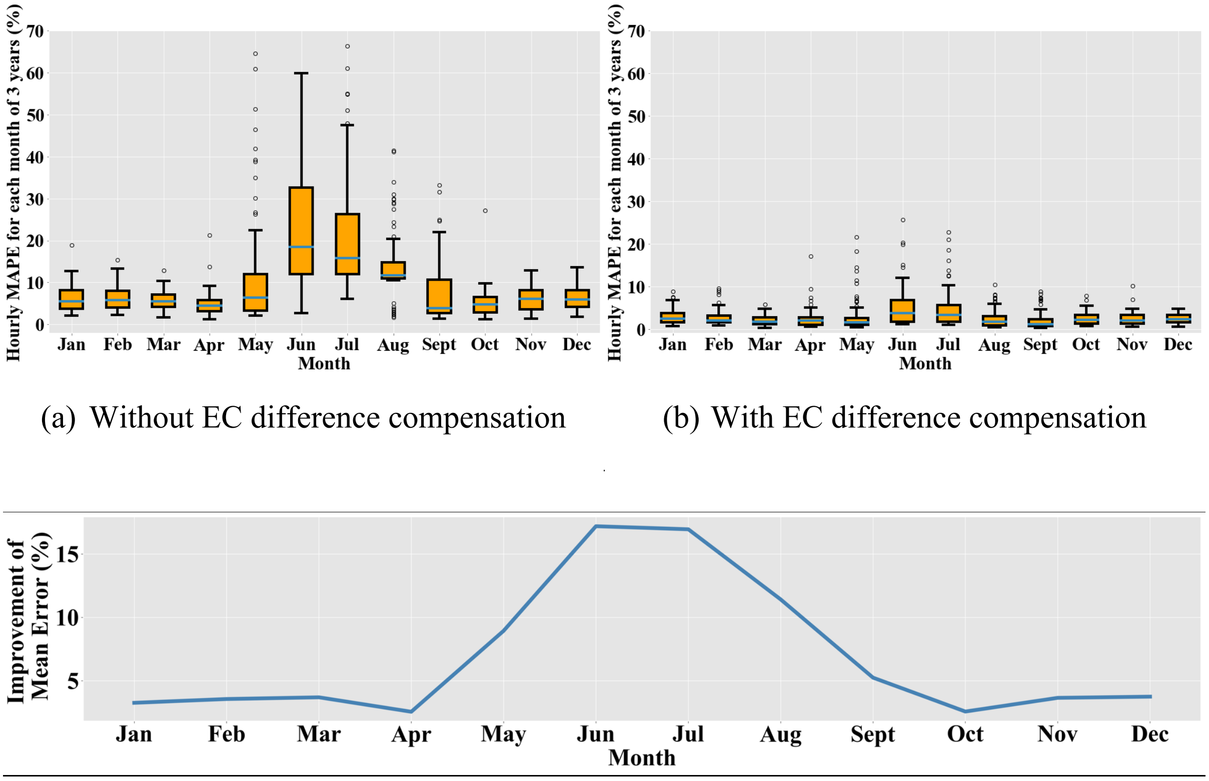
Methods	**Performance improvement**					
	Compensated, LSTM, *c*=2		**Compensated, LSTM, *c*=3**		**Compensated, LSTM, *c*=4**	
	RMSE	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**
Method in [14] (RMSE: 2.61 kW, MAPE: 8.26%)	**1.17 kW**	**5.06%**	0.52 kW	1.35%	0.57 kW	1.64%
Method in [15] (RMSE: 2.35 kW, MAPE: 9.10%)	**1.10 kW**	**5.87%**	0.45 kW	2.16%	0.50 kW	2.45%

Table 7 shows the reduction ranges of the proposed model built using LSTM with compensation of average hourly RMSE and MAPE for 3 years are between 0.52 kW-1.17 kW and 1.35%-5.06% respectively, as compared to method in [14]. The reduction ranges of average hourly RMSE and MAPE for 3 years are between 0.45 kW-1.10 kW and 2.16%-2.45% respectively, as compared to method in [15].

5.4. Performance analysis of the proposed disaggregation method

5.4.1. Performance analysis of the semi-synthetic datasets

To further analyze the performance of the proposed method when using meteorological data that well-matched the PV system, the hourly MAPE for each month modeled utilizing semi-synthetic data of PVPG with and without EC difference compensation is shown in Fig. 12.



(a) Without EC difference compensation

(b) With EC difference compensation

(c) Mean error improvement with EC difference compensation of each month

Fig. 12. Hourly MAPE for each month comparison of PVPG disaggregation with and without EC difference compensation for 3 years

Comparing Fig. 12(a) with Fig. 12(b), after having EC difference compensation, both the accuracy and stability of disaggregation have been greatly improved. Before the compensation, the largest median value of hourly MAPE and the largest hourly MAPE are concentrated from May to September, however the median and largest value of hourly MAPE are greatly reduced in these months after the compensation. Fig. 12(c) further shows after the compensation, the average hourly MAPE reduction of each month is relatively obvious from May to September, while the error reduction in other months is relatively small.

To explain this change, average daily temperature for each month of this region is displayed in Fig. 13 and the maximal information coefficient (MIC) [12] is used to calculate the correlation between average daily temperature and average daily EC for each month in the region.

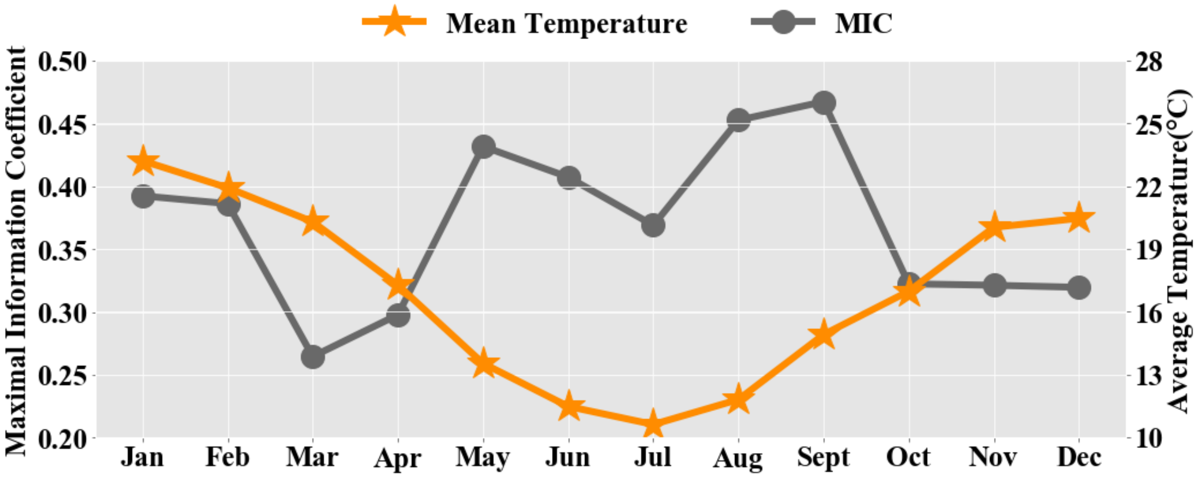
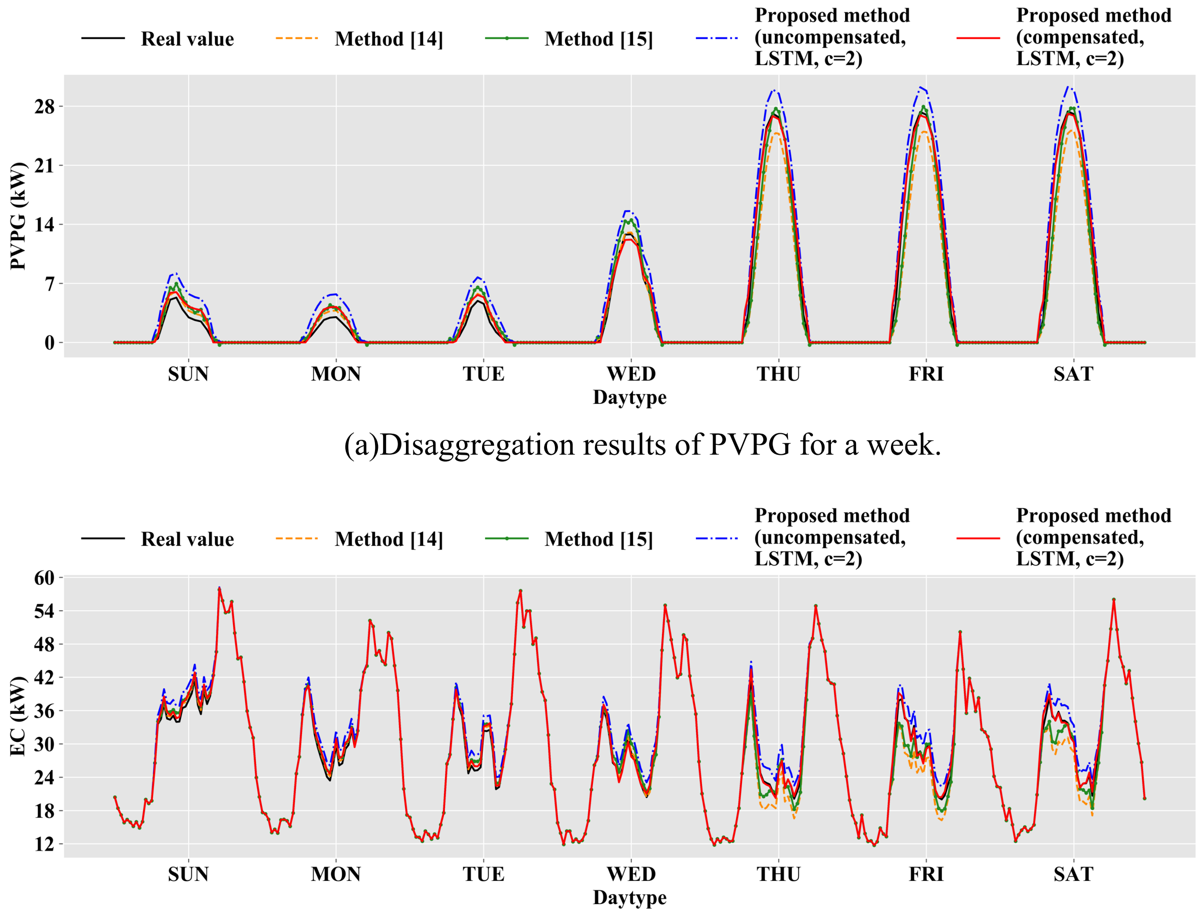


Fig. 13. Average temperature and the MIC between average temperature and average EC.

Fig. 13 shows MIC is high from May to September, which means that in these months, the EC behaviors are largely dependent on the temperature. This can also explain why the error reduction in these months is so obvious after the EC difference compensation. In Australia, the duration of winter is usually defined from early June to end of August. During this period, the extensive use of electrical equipment for heating has led to a larger proportion of temperature-sensitive loads. After the compensation, the errors from January to April and November are relatively large in the whole year due to these months belong to the coverage of Australia's spring and autumn. From Fig. 13, we can see that the average temperature is between 15 °C and 20 °C, which will lead to less temperature-sensitive load and cause less temperature-driven changing of EC. If the temperature-based compensation is performed at this situation, it is likely to cause the correction to be inconspicuous, or even to incorrectly explain the reason for the change in EC. Fig. 14 provides the PVPG and EC disaggregation results of a week in June. This shows that the curves estimated by the proposed method with compensation are closer to the real value in both amplitude and shape. In Fig. 14, the estimated results from Sunday to Wednesday show that the disaggregation method with compensation proposed can well fit the shape of PVPG even under non-clear sky conditions. Since the results of EC disaggregation is obtained by adding net load and disaggregated PVPG, Fig. 14(b) shows that the disaggregated EC also has a higher accuracy when the results of the PVPG disaggregation are accurate.



(a)Disaggregation results of PVPG for a week.

(b)Disaggregation results of EC for a week.

Fig. 14. Disaggregation results comparison of different methods for a week.

Disaggregation results with and without EC difference compensation in different sub-regions for 3 years are shown in Table 8.

Table 8

Disaggregation results with and without EC difference compensation in different sub-regions

Sub-region	Without EC difference compensation				With EC difference compensation			
	MLP		LSTM		MLP		LSTM	
	RMSE (kW)	MAPE (%)	RMSE (kW)	MAPE (%)	RMSE (kW)	MAPE (%)	RMSE (kW)	MAPE (%)
1	2.52	13.73	1.86	10.88	1.56	9.37	**1.99**	**8.49**
2	1.96	13.03	1.43	10.55	0.89	5.08	**0.59**	**3.20**
3	0.94	14.63	0.62	11.25	0.91	9.36	**0.59**	**7.02**
4	1.86	13.99	0.94	8.79	1.02	9.21	**0.93**	**6.08**
5	2.91	16.43	2.00	11.48	1.43	8.28	**1.33**	**7.82**

Considering the 5 sub-regions, Table 8 shows that the disaggregation method with EC difference compensation provides a higher accuracy as compared to the method without EC difference compensation. From the disaggregation results with EC difference compensation, Sub regions 3 and 4 have relatively low average hourly RMSE, and Sub-regions 1, 2 and 5 have relatively high average hourly RMSE, which is mainly affected by the aggregation level. The aggregation degree of the former regions is lower than the latter one, so the cardinal number of net load is relatively small. For all 5 sub-regions, LSTM has a better decomposition accuracy as compared to MLP, which also illustrates the advantages of LSTM in disaggregating the experimental data set.

5.4.2. Performance analysis of the raw datasets

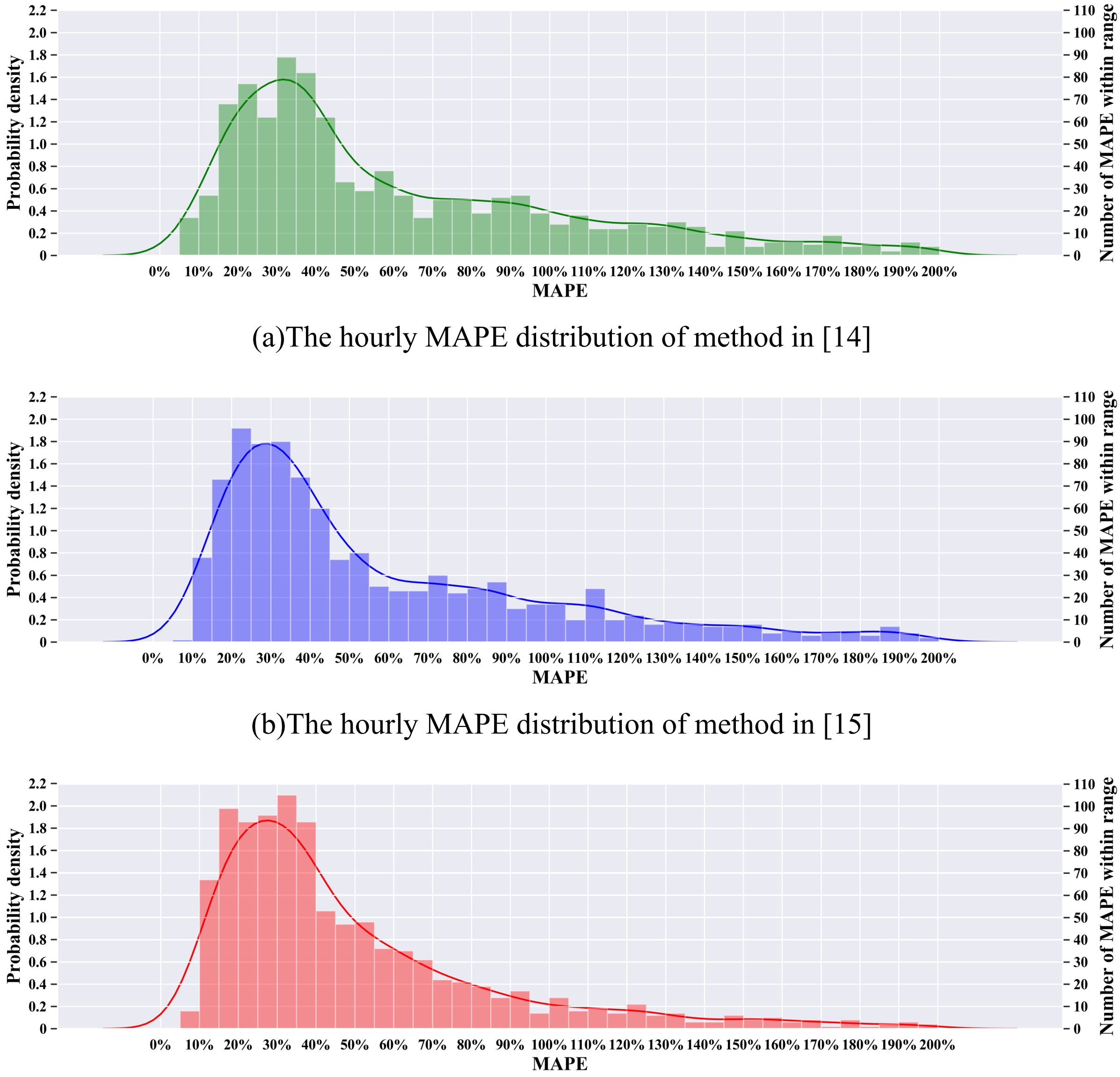
In reality, due to the lack of sensing devices, it is not easy to obtain accurate meteorology data about the target customers' geographical location. The output of PV equipment is extremely sensitive to the cloud cover state, the difference in irradiation between the meteorological data collection point and the target location introduces a large amount of measurement error, which causes a decrease in the accuracy of the disaggregation algorithm. In order to analyze the robustness of the proposed disaggregation algorithm, the raw PVPG data of Ausgrid and the meteorological data from the sub-regional central area are used for model building. For simplicity, proposed method (compensated, LSTM, $c = 2$) will be presented.

The hourly MAPE distributions for 3 years in Sub-region 2 are presented in Table 9.

Table 9
The hourly MAPE distributions for 3 years in Sub-region 2

Methods	Intervals of MAPE					
	$[0\%, 100\%)$	$[100\%, 200\%)$	$[200\%, 300\%)$	$[300\%, 400\%)$	$[400\%, 500\%)$	$[500\%, +\infty)$
Method in [14]	769	177	54	29	16	51
Method in [15]	804	157	49	28	13	45
Proposed method	911	103	31	12	8	31

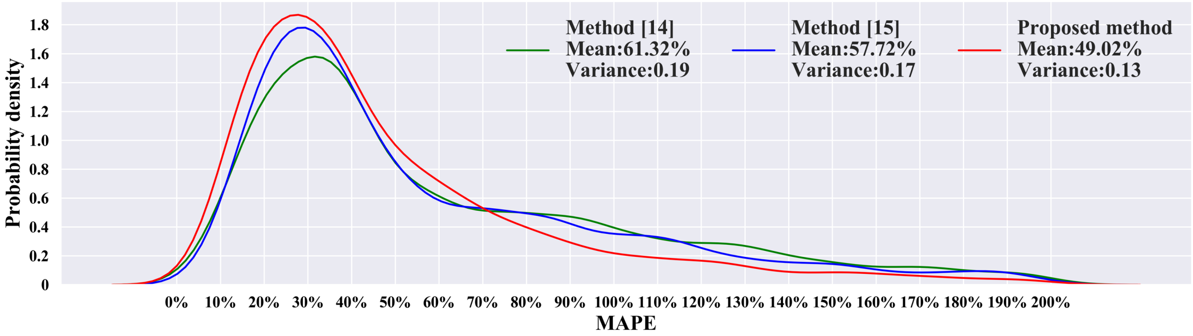
Due to spatial differences, the collected meteorological data may not truly reflect the actual irradiance received by the PV systems of the target user, that is, the site for data collection may be different from that of PV equipment installation. In this case more than 500% PVPG disaggregation MAPE can occur. Table 9 shows that among the disaggregation results for a total of 1096 days, the percentage of days with MAPE vary from 0-200% for method in [14], method in [15] and proposed method are 86.31%, 87.68% and 92.52%, respectively. Detailed MAPE distribution results are shown in Fig. 15 by histograms and the Gaussian kernel density estimation curves.



(a)The hourly MAPE distribution of method in [14]

(b)The hourly MAPE distribution of method in [15]

(c)The hourly MAPE distribution of proposed method



(d)Probability density curves of different methods

Fig. 15 The Histograms and the probability density curves of hourly MAPE of different methods

Fig. 15(d) shows that the peak probability density of each of the three methods occurs at MAPE of 31.86%, 28.83% and 27.06%. The MAPE probability density curve of the proposed method intersects with method in [14] and method in [15] around MAPE of 70%. On the left-hand side of the intersection point, the proposed method has a higher probability density. The mean values of method in [14], method in [15] and proposed method are 61.32%, 57.72% and 49.02% respectively.

The inaccuracy of the disaggregation for small actual PVPG will greatly increase the absolute percentage error, for example, when the actual PVPG is 0.1 kW and the disaggregation result is 0.2 kW, the absolute percentage error is 100%. The above case will contribute much more to the average hourly MAPE, which will render the average hourly MAPE metrics statistically meaningless. Therefore, we use average hourly RMSE of each month for the metric. The average hourly RMSE of each month for 2012 is shown in Table 10.

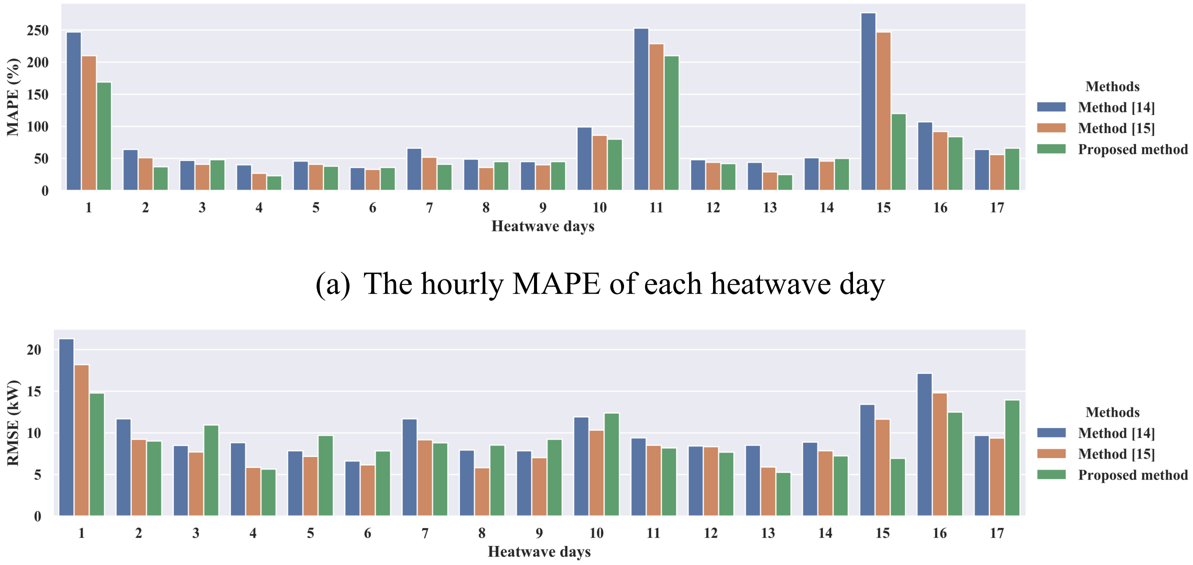
Table 10

The average hourly RMSE of each month for 2012

Methods	RMSE (kW)											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
Method in [14]	13.42	12.30	11.07	5.45	**3.94**	**3.97**	4.50	4.68	**5.21**	9.12	10.17	12.52
Method in [15]	11.80	11.08	10.16	5.30	4.09	3.98	4.65	4.78	5.43	8.39	9.14	10.91
Proposed method	**9.66**	**9.19**	**9.03**	**5.27**	4.13	4.16	**4.20**	**4.53**	6.00	**8.36**	**8.90**	**8.88**

It can be observed from Table 10 that the proposed method achieves the lowest average hourly RMSE in all months except May, June and September. Therefore, the statistical results from Fig. 15 and Table 10 both indicate that the proposed algorithm has better robustness in the raw data set.

Heatwave occurs often in Australia, and the increase in temperature can significantly reduce the photovoltaic conversion efficiency of PV systems. In order to investigate the performance of the proposed disaggregation method under extreme hot weather, 17 days with the hottest maximum temperature were selected from July 1, 2010 to June 30, 2013 for the results analysis. The heatwave dates are described in Appendix A. The hourly RMSE and MAPE of each heat-wave day are shown in Fig. 16.



(a) The hourly MAPE of each heatwave day

(b) The hourly RMSE of each heatwave day

Fig. 16 The hourly RMSE and MAPE of 17 heatwave days

Results of Fig. 16 denote that the proposed method has the minimum MAPE value for 12 out of the 17 heatwave days, and has the minimum RMSE value for 11 out of the 17 heatwave days among the three disaggregation methods. This illustrates the effectiveness of the proposed disaggregation method in extremely hot days.

6. Conclusion and future work

In practice, it is difficult to obtain complete and independent PVPG data which is attributed by the following four points: (1) When considering the economy, rooftop PV panels are installed BtM and PVPG data will not be recorded separately. (2) There may be equipment problems with PVPG meter, resulting the complete or accurate PVPG data cannot be recorded. (3) Due to the timeliness of information, the PV panels installation information recorded by the utility may not match the actual situation. (4) Users may illegally install PV panels without the permission of energy system operator. Incorrect or vague information will increase the difficulty of assessing regional PV potential, escalating the risk of connection to the main grid, and reducing the stability of the power system.

To solve these practical problems, this paper presents a novel unsupervised data-driven BtM PVPG disaggregation method. This disaggregation method does not need to set up PV system proxy and no reliance on PV physical model consumption. Therefore, it has strong practicability and can still guarantee the disaggregation accuracy when facing the degradation of the PV arrays.

The most unique innovation of this work is that based on the conversion characteristics of PV panels, the PVPG sensitivity model can be converted into a PVPG estimation model, which can perform the net load disaggregation when only historical net load data and historical weather data are accessible. In order to refine the model, clustering is used to divide EC behaviors before disaggregation, and the accuracy of the PVPG disaggregation model is further improved through EC difference compensation. The comparison of results with two state-of-the-art unsupervised disaggregation models prove that the proposed method has higher disaggregation accuracy. Net load datasets in Ausgrid at residential level are employed in the case study. Disaggregation results of semi-synthetic datasets of 5 separated sub-regions have shown the effective and satisfactory. The disaggregation results of the raw datasets with meteorological data from heterogeneous collection sources and heatwave days have shown the robustness and practicality of the proposed method.

This paper explored a disaggregation method for net load as a consequence of BtM PVPG supply and proposed a new data-driven BtM PVPG disaggregation architecture. However, only a single type of distributed energy system is considered and the model is still flawed. Hence, the authors will focus on more complex BtM scenes and broader applications. The future work includes: 1) optimizing the model architecture to obtain higher disaggregation accuracy, 2) exploring more types of net load disaggregation methods, such as considering wind power or biogas, and 3) applying BtM disaggregation results to scenarios more broadly, such as energy storage and net load forecasting.

Acknowledgements

The work is sponsored by the Department of Finance and Education of Guangdong Province 2016 [202]: Key Discipline Construction Program, China; the Education Department of Guangdong Province: New and Integrated Energy System Theory and Technology Research Group [Project Number 2016KCXTD022]; National Natural Science Foundation of China (51907031) and Brunel University London BRIEF Funding, UK.

Appendix A. The detailed parameters of the experimental dataset

The 1096 days from July 1, 2010 to June 30, 2013 are numbered from1 to 1096. The date remarked as heatwave is 926, 209, 932, 883, 215, 219, 922, 925, 214, 210, 884, 216, 915, 208, 220, 903, 233 (listed in descending order of the highest daily temperature).

Table A

The ID of the customers with anomalous general load data and controllable data, and the ID of the customers in remote areas

	ID
Customers in remote areas, with anomalous general load data and controllable data	2, 9, 27, 34, 57, 68, 85, 95, 104, 121, 143, 146, 150, 152, 176, 187, 191, 212, 221, 229, 239, 248, 249, 260, 265, 272, 273, 284, 287, 289, 293, 294

Table B

The ID of the sub-regions division of the 269 customers

Sub-regions	ID
1	4, 6, 11, 12, 16, 17, 31, 32, 36, 40, 43, 44, 45, 48, 53, 55, 56, 60, 62, 66, 71, 78, 79, 80, 92, 96, 98, 99, 105, 113, 116, 118, 120, 126, 128, 129, 131, 135, 139, 140, 142, 156, 159, 162, 164, 172, 173, 181, 182, 192, 193, 194, 199, 205, 213, 217, 219, 220, 226, 233, 234, 237, 242, 244, 251, 257, 262, 264, 268, 275, 277, 278, 279, 282, 288, 291, 298
2	1, 8, 18, 32, 46, 63, 65, 74, 76, 81, 82, 83, 89, 91, 94, 97, 100, 102, 109, 112, 114, 115, 117, 123, 132, 134, 136, 141, 147, 149, 154, 161, 166, 174, 179, 180, 183, 185, 190, 195, 197, 198, 204, 208, 209, 210, 216, 228, 235, 236, 241, 243, 247, 250, 254, 255, 258, 274, 280, 295, 296, 299, 300
3	5, 13, 21, 28, 50, 54, 58, 61, 69, 70, 72, 75, 86, 90, 127, 158, 165, 167, 178, 224, 225, 245, 246, 266, 271, 276, 286, 292, 297
4	3, 7, 10, 14, 15, 19, 20, 23, 29, 30, 37, 38, 39, 42, 49, 64, 67, 84, 101, 106, 130, 137, 145, 155, 160, 168, 169, 171, 177, 184, 186, 189, 196, 202, 206, 215, 218, 223, 227, 232, 238, 267, 270, 283
5	22, 24, 25, 26, 33, 35, 41, 47, 51, 52, 59, 73, 77, 87, 88, 93, 103, 107, 108, 110, 111, 119, 122, 124, 125, 133, 138, 144, 148, 151, 153, 157, 163, 170, 175, 188, 200, 201, 203, 207, 211, 214, 222, 230, 231, 240, 252, 253, 256, 259, 261, 263, 269, 281, 285, 290

Table C

Coordinates of the measured meteorological data from the central area in the sub-regions

Sub-regions	Coordinates
1	33°56'S, 151°41'E
2	33°12'S, 151°27'E
3	33°23'S, 151°25'E
4	33°40'S, 151°09'E
5	33°54'S, 151°07'E

Appendix B. The hyper-parameters of the artificial neural network

When building the PVPG and EC sensitivity model, Adam algorithm is adopted with a mini-batch size of 256 and learning rate 0.0001. All the disaggregation results are obtained from the average values of 30 experiments.

Table D

Hyper-parameters of the artificial neural network

Models	Layers	Iterations
EC-MLP	$[200,20]$[a]	20000
EC-LSTM	$[200,20]$	20000
PVPG-MLP	$[200,100]$	30000
PVPG-LSTM	$[200,100]$	30000

[a]: $[\bullet,\bullet]$ denotes that there are two hidden layers and each layer contains $\bullet$ neurons.

REFERENCES

[1] Lai CS, Locatelli G, Pimm A, Wu X, Lai LL. A review on long-term electrical power system modeling with energy storage. J Clean Prod 2021;280:124298. https://doi.org/10.1016/j.jclepro.2020.124298.

[2] Lai CS, Jia Y, Lai LL, Xu Z, McCulloch MD, Wong KP. A comprehensive review on large-scale photovoltaic system with applications of electrical energy storage. Renew Sustain Energy Rev 2017;78:439–51. https://doi.org/10.1016/j.rser.2017.04.078.

[3] Lai CS, McCulloch MD. Levelized cost of electricity for solar photovoltaic and electrical energy storage. Appl Energy 2017;190:191–203. https://doi.org/10.1016/j.apenergy.2016.12.153.

[4] Xu X, Li J, Xu Y, Xu Z, Lai CS. A Two-stage game-Theoretic method for residential pv panels planning considering energy sharing mechanism. IEEE Trans Power Syst 2020;35:3562–73. https://doi.org/10.1109/TPWRS.2020.2985765.

[5] Australian Energy Market Operator. Draft 2020 Integrated System Plan 2019:1–83.

[6] Mason K, Reno MJ, Blakely L, Vejdan S, Grijalva S. A deep neural network approach for behind-the-meter residential PV size, tilt and azimuth estimation. Sol Energy 2020;196:260–9. https://doi.org/10.1016/j.solener.2019.11.100.

[7] Kabir F, Yu N, Yao W, Yang R, Zhang Y. Joint estimation of behind-the-meter solar generation in a community. IEEE Trans Sustain Energy 2021;12:682–94. https://doi.org/10.1109/TSTE.2020.3016896.

[8] Gordon JM, Fasquelle T, Nadal E, Vossier A. Providing large-scale electricity demand with photovoltaics and molten-salt storage. Renew Sustain Energy Rev 2021;135:110261. https://doi.org/10.1016/j.rser.2020.110261.

[9] Shaker H, Zareipour H, Wood D. A data-driven approach for estimating the power generation of invisible solar sites. IEEE Trans Smart Grid 2016;7:2466–76. https://doi.org/10.1109/TSG.2015.2502140.

[10] Bu F, Dehghanpour K, Yuan Y, Wang Z, Zhang Y. A Data-Driven Game-Theoretic Approach for Behind-the-Meter PV Generation Disaggregation. IEEE Trans Power Syst 2020. https://doi.org/10.1109/TPWRS.2020.2966732.

[11] Li K, Wang F, Mi Z, Fotuhi-Firuzabad M, Duić N, Wang T. Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation. Appl Energy 2019. https://doi.org/10.1016/j.apenergy.2019.113595.

[12] Stainsby W, Zimmerle D, Duggan GP. A method to estimate residential PV generation from net-metered load data and system install date. Appl Energy 2020;267:114895. https://doi.org/10.1016/j.apenergy.2020.114895.

[13] Chen D, Irwin D. SunDance: Black-box behind-the-meter solar disaggregation. E-Energy 2017 - Proc 8th Int Conf Futur Energy Syst 2017:45–55. https://doi.org/10.1145/3077839.3077848.

[14] Wang Y, Zhang N, Chen Q, Kirschen DS, Li P, Xia Q. Data-driven probabilistic net load forecasting with high penetration of behind-the-meter PV. IEEE Trans Power Syst 2018;33:3255–64. https://doi.org/10.1109/TPWRS.2017.2762599.

[15] Sossan F, Nespoli L, Medici V, Paolone M. Unsupervised disaggregation of photovoltaic production from composite power flow measurements of heterogeneous prosumers. ArXiv 2017;14:3904–13.

[16] Ausgrid n.d. https://www.ausgrid.com.au/SearchResults?cs =data (accessed January 5, 2021).

[17] Orme S, Swansson J. Implications of extreme weather for the Australian National Electricity Market:

historical analysis and 2019 extreme heatwave scenario. 2014.

[18] Ratnam EL, Weller SR, Kellett CM, Murray AT. Residential load and rooftop PV generation: an Australian distribution network dataset. Int J Sustain Energy 2017;36:787–806. https://doi.org/10.1080/14786451.2015.1100196.

[19] Pfenninger S, Staffell I. Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. Energy 2016;114:1251–65. https://doi.org/10.1016/j.energy.2016.08.060.

[20] Renewables.ninja n.d. https://www.renewables.ninja/ (accessed January 5, 2021).

[21] Google Earth n.d. https://www.google.com/intl/zhCN/earth/ (accessed January 5, 2021).

[22] Hoke A, Butler R, Hambrick J, Kroposki B. Steady-state analysis of maximum photovoltaic penetration levels on typical distribution feeders. IEEE Trans Sustain Energy 2013;4:350–7. https://doi.org/10.1109/TSTE.2012.2225115.

[23] Keller JM, Gray MR. A fuzzy K-nearest neighbor algorithm. IEEE Trans Syst Man Cybern 1985;SMC-15:580–5. https://doi.org/10.1109/TSMC.1985.6313426.

[24] Lai CS, Yang Y, Pan K, Zhang J, H.L. Yuan HL, Ng WWY, et al. Multi-view neural network ensemble for short and mid-term load forecasting. IEEE Trans Power Syst 2020:1–1. https://doi.org/10.1109/TPWRS.2020.3042389.

[25] Wang D, Wu R, Li X, Lai CS, Wu X, Wei J, et al. Two-stage optimal scheduling of air conditioning resources with high photovoltaic penetrations. J Clean Prod 2019;241. https://doi.org/10.1016/j.jclepro.2019.118407.

[26] Chi F, Xu L, Pan J, Wang R, Tao Y, Guo Y, et al. Prediction of the total day-round thermal load for residential buildings at various scales based on weather forecast data. Appl Energy 2020;280:116002. https://doi.org/10.1016/j.apenergy.2020.116002.

[27] Lai CS, Jia Y, McCulloch MD, Xu Z. Daily clearness index profiles cluster analysis for photovoltaic system. IEEE Trans Ind Informatics 2017;13:2322–32. https://doi.org/10.1109/TII.2017.2683519.

[28] Lusis P, Khalilpour KR, Andrew L, Liebman A. Short-term residential load forecasting: Impact of calendar effects and forecast granularity. Appl Energy 2017;205:654–69. https://doi.org/10.1016/j.apenergy.2017.07.114.

[29] Teeraratkul T, O’Neill D, Lall S. Shape-based approach to household electric load curve clustering and prediction. IEEE Trans Smart Grid 2018. https://doi.org/10.1109/TSG.2017.2683461.

[30] Tang R, Yildiz B, Leong PHW, Vassallo A, Dore J. Residential battery sizing model using net meter energy data clustering. Appl Energy 2019;251:113324. https://doi.org/10.1016/j.apenergy.2019.113324.