Sensitivity analysis of distributed photovoltaic system capacity estimation based on artificial neural network

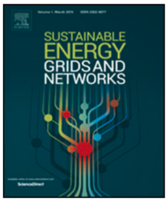
Tang, Lingxi; Ashtine, Masao; Hua, Weiqi; C.H. Wallom, David

DOI:
10.1016/j.segan.2024.101396

Contents lists available at ScienceDirect

Sustainable Energy, Grids and Networks

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



Sensitivity analysis of distributed photovoltaic system capacity estimation based on artificial neural network

Lingxi Tang, Masaō Ashtine, Weiqi Hua, David C.H. Wallom [*]

Department of Engineering Science, University of Oxford, Oxford OX1 3QG, UK

ARTICLE INFO

Keywords:
Data analytics
Feature extraction
Feeder load data
Machine learning
Solar photovoltaic
Capacity estimation

ABSTRACT

Residential solar photovoltaic (PV) system installations are expected to continue increasing due to their growing cost competitiveness and supportive government policies. However, excessive installations of unknown behind-the-meter solar panels present a challenge for accurate load prediction and reliable operations of power networks. To address such growing concerns of distribution network operators (DNOs), this research proposes a novel model for distributed PV system capacity estimations. Innovative extracted features from 24-hour substation net load curves were fed into a deep neural network to estimate the PV capacity linked to the substation feeder. A comprehensive study into the sensitivity of the model's accuracy to specific temporal scales of data collection, number of households served by a substation, and proportion of PV-equipped properties was conducted. This study revealed that a model developed to be used exclusively in summer achieved a 18.1% decrease in estimation root mean squared error (RMSE) compared to an all-year model, whilst using only a third of the training data amount. Similarly, compared to an all-year model, RMSE decreased by 26.9% when only data from Mondays to Thursdays were used to train and test the model. Also, for the all-year model, the most accurate estimations occur when 20% to 80% of households have PV systems installed and estimation percentage error tend to remain constant at around 10% when more than 20% of households have PV systems installed. A machine learning-ready dataset of substations with known PV capacity and experiment results are both useful to inform DNOs on the potential of the proposed method in reducing grid operation costs.

1. Introduction

Regulatory supports and major cost reductions have encouraged the growth of distributed photovoltaic systems (DPVS) globally. Global PV capacity has increased rapidly in recent years, with solar energy capacity more than double from 493 GW in 2018 to over 1060 GW in 2022 [1]. Furthermore, global installed PV capacity is expected to grow at an increasing rate, with annual additions of 400 GW in 2024, up to over 500 GW in 2028 [2]. In the UK, total small-scale DPVS[1] deployment has grown from about 1.4 GW to 3.5 GW over the last decade and accounts for 23.1% of all solar PV installations as of January 2024 [3]. A resulting issue is a large number of DPVSs being unknown or incorrectly registered with distribution network operators (DNOs) [4,5], which leads to a variety of techno-economic issues. For example, high penetrations of unknown DPVS would result in a multitude of technical issues, e.g., over-voltage [6], and reverse power flow [7], to potentially threaten the reliability and security of supply of power networks [8]. Also, these unknown DPVS reduce the accuracy of load forecasting and estimation for demand reduction potentials [5,9].

To avoid costs of manually checking and monitoring for DPVS installations, developing automatic approaches for the distributed PV system capacity estimation (DPVSCE) has been the focus of recent research practices. DPVSCE approaches can be categorised as either based on satellite imagery data or electricity load data [10]. Three recent major studies have looked into using satellite images to extract the available amount of rooftop spaces, which was used to infer potential PV installations. Zhong et al. [11] and Krapf et al. [12] both extracted rooftop area from satellite aerial images, using image semantic segmentation, to determine potential solar power generated and profits respectively. Ren et al. [13] combined rooftop area obtained from satellite imagery, with solar irradiance profile obtained from 3D Geographic Information System, to calculate solar energy potential on individual buildings.

Although research have demonstrated the feasibility of using satellite imagery to determine solar power potential, satellite imaged-based DPVSCE present four primary challenges from the perspective of DNOs:

[*] Corresponding author.
E-mail addresses: lingxi.tang@gess.ethz.ch (L. Tang), masao.ashtine@carbontrust.com (M. Ashtine), w.hua@bham.ac.uk (W. Hua), david.wallom@eng.ox.ac.uk (D.C.H. Wallom).

[1] Less than or equal to 4 kW capacity.

https://doi.org/10.1016/j.segan.2024.101396
Received 12 December 2023; Received in revised form 20 March 2024; Accepted 16 April 2024
Available online 3 May 2024

(1) The focus of identifying potential PV installations, which does not reveal immediate implications of installed unknown DPVS; *(2)* DNOs are unable to directly access the satellite imagery data without a third party, which incurs huge data purchasing costs and administrative delays; *(3)* Using satellite imagery data assume a constant relationship between the physical size of solar panels and their electricity generation capacities. However, the capacity per unit area of a solar panel can vary greatly among different arrays due to differences in the efficiency; *(4)* Satellite imagery data has limited information on the substation feeder to which a solar panel connects.

With feeder-level data being more accessible for DNOs [4], research based on utilising electricity load data has risen in popularity and was featured in multiple recent studies to perform net load PV disaggregation. Pan et al. [14] compiled a list of recent PV disaggregation studies in their paper and subsequently proposed a 3-stage framework to obtain PV generation data from net load data. They used machine learning (ML) to develop models which approximate PV generation based on irradiation, temperature and net load consumption data. Wang et al. [15] proposed a two-stage approach for the PV detection and estimation, in which the first stage used support vector classification to detect the presence of the DPVS from a household net-load curve, and the second stage utilised long short term memory [16] with specific extracted features to obtain the amount of generated PV power. Liu et al. [17] proposed a self-supervised learning method where a separate set of PV generation measurements were used to generate pseudo labels for unlabelled net load data, before using the newly labelled data to train a estimation model. Zhang et al. [18] proposed a probabilistic model based on the multi-quantile recurrent neural networks to disaggregate PV generation and separate demand into subcomponents. These works demonstrate the utility of using electricity consumption data in separating electricity generation and consumption within net-load curves.

Two major gaps remain in existing research: *(1)* They focus solely on PV generation estimation, which does not provide information on PV capacity estimation or the maximum potential PV generation; *(2)* Existing research requires multiple types of input data, e.g. solar irradiance, load profile of individual households without PV or with known PV capacity, and feeder net-load data, which are not always obtainable in practical applications.

This paper fills the gap of existing research by offering the following contributions:

- An automatic DPVSCE model is proposed through developing a novel data-driven approach of feature extraction and ML. The proposed DPVSCE model only requires the feeder-level net-load data, which is cost-effective and scalable for various distribution networks, and focuses on forecasting installed PV capacity. An in-depth analysis of the model's accuracy in response to changing variables is also conducted, identifying key factors affecting estimation accuracy.
- A new set of demand behaviour groupings within a week is introduced and analysis revealed clear distinctions in electricity consumption behaviour between 3 time groups within a week.
- A dataset containing a total of 25,748 samples of input features extracted from 24-hour net load time-series data and their corresponding output label of installed PV capacity is made available and can be accessed via the GitHub repository relevant to this paper.[2] This dataset is ready for ML training.

The rest of this paper is organised as follows: Section 2 explains the proposed DPVSCE model, substantiated by electricity behaviour analysis based on a new set of intra-week groupings. The simulation methods used for hyper-parameter optimisation, data preparation and sensitivity analysis are discussed in Section 3. Section 4 presents and discusses the results obtained from the simulations conducted. Section 5 concludes this paper, by highlighting key findings and recommendations for DNOs, and listing areas for further research. For convenience, Table 1 represents the nomenclature of this paper.

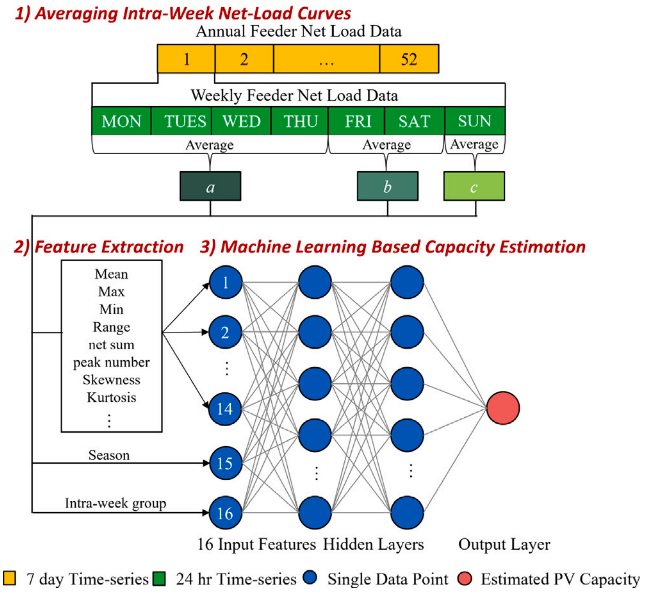


Fig. 1. Overall framework of the proposed DPVSCE model. This framework consists of three major steps: *(1)* averaging intra-week net-load curves, *(2)* feature extraction, and *(3)* ML-based capacity estimation.

2. Model design

This section introduces the proposed DPVSCE model, by first presenting the model's overall framework before explaining the feature extraction and ML techniques featured in the model.

2.1. Overall framework

The rationale behind the proposed model is that the installed DPVS capacity directly determines PV generation output, and subsequently affects the net-load. This relationship is illustrated by the following two equations:

$$p_t^{\mathrm{pv}} = c \cdot \eta_t, \tag{1}$$

$$p_t^{\mathrm{net}} = p_t^{\mathrm{c}} - p_t^{\mathrm{pv}}, \tag{2}$$

where p_t^{pv} is the total PV output of all solar panels connected to a substation at the time step t, c is the installed capacity of all solar panels connected to a substation, η_t is the generation efficiency of solar panels connected to a substation at the time step t, p_t^{net} is the net power load of a substation at the time step t, and p_t^{c} is the power consumption of a substation at the time step t. Positive values of p_t^{net} indicate power consumption, whereas negative values of p_t^{net} indicate power generation. Therefore, the installed DPVS capacity c can theoretically be inferred from the feeder-level net-load p_t^{net}, forming the basis of DPVSCE based on net-load data.

Fig. 1 presents the overall framework of the proposed DPVSCE model which takes a 24-hour feeder-level net-load curve as the input, extracts the relevant features, and outputs the estimated installed DPVS capacity associated with the substation feeder. This framework consists of three major steps: *(1)* averaging intra-week net-load curves, *(2)* feature extraction, and *(3)* ML-based capacity estimation.

[2] https://github.com/LingxiTang/ML-DPVSCE-Paper

Table 1
This table represents the nomenclature of this paper.

Nomenclature			
PV	Photovoltaic	AE	Absolute Error
DPVS	Distributed photovoltaic system	MAE	Mean Absolute Error
DPVSCE	Distributed photovoltaic system capacity estimation	MAPE	Mean Absolute Percentage Error
DNO	Distribution network operator	RMSE	Root Mean Squared Error
PVGIS	Photovoltaic Geographical Information System	ML	Machine learning
IFEEL	Interpretable Feature Extraction of Electricity Loads	ANN	Artificial neural network
EV	Electric Vehicles	SAX	Symbolic aggregate approximation

From Eqs. (1)–(2), the following equation is obtained:

$$c = \frac{p_t^{\mathrm{c}} - p_t^{\mathrm{net}}}{\eta_t}, \tag{3}$$

hence, the installed DPVS capacity c is obtainable from the net-load data p_t^{net} if the power consumption p_t^{c} and generation efficiency η_t is known.

Step 1 extracts known characteristics of p_t^{c} and η_t,by sorting input net-load curves into distinct intra-week groups with similar consumption patterns and then averaging data within each group. A common grouping found in relevant literature would be weekdays and weekends [19,20]. In other cases, Saturdays, Sundays and Mondays have been separated as distinct groups [21,22]. In this paper, a unique grouping will be tested: *(a)* Monday to Thursday, *(b)* Friday to Saturday, and *(c)* Sunday. The reason for grouping Monday to Thursday together is that they are typical working days which have similar load patterns. Friday and Saturday are grouped together due to these days being ‘out nights’, characterised by the general behaviours of consumers returning home late on these days, reducing/delaying the evening peak in electricity consumption. Sunday is subsequently classified into a distinct group, separate from the first two groups. The identified intra-week groups act as features which provide profile classifications for household power consumption p_t^{c}. For the generation efficiency η_t, the main source of variation is the amount of solar irradiation, which is largely dependent on the season, i.e. summer, winter or transitional.[3] Thus, the “Season” feature (see Fig. 1) provides information on η_t.

Step 2 extracts key features from the input averaged net-load curve. A total of 16 features were utilised as inputs for our proposed model. Out of these input features, 14 were based on the Interpretable Feature Extraction of Electricity Loads (IFEEL) package, a tool designed to extract interpretable features from daily electricity load curves and had been proven effective at detecting PV installations from intraday load curves [24]. Two other input features were created based on the timestamp of the input 24-hour net-load curve, i.e., intra-week group and season. The season feature serves to distinguish between season-specific household load consumption behaviours. For instance, electricity demand, in the UK, is generally higher in winter than in summer due to residential heating requirements versus a lack of cooling demand, and a steeper evening surge is usually presented during winter, as compared to summer [25].

Step 3 maps the extracted features to the estimated DPVS capacity through an artificial neural network (ANN), which will be presented in Section 2.3.

2.2. Feature extraction

The feature extraction process in the model has three objectives: *(1)* dimensionality reduction, i.e., removing the dimension of time, by converting the time-series load curve into interpretable feature values, *(2)* creating typical features of consumption patterns to be learnt by the ANN, and *(3)* allowing the proposed model to be applicable to input data of any time resolution due to the removal of the time dimension. The IFEEL package [24] extracts 14 out of 16 input features for the proposed DPVSCE model, which are listed in Table 2. Fig. 2 shows how the input features describe curve characteristics.

Table 2
Explanation of input features from IFEEL.

No.	Input feature	No.	Input feature
1	Mean of 24-hour load curve (kW)	8	Sum of net-loads during non -business hours (kW)
2	Standard deviation of 24-hour load curve (kW)	9	Skewness of 24-hour load curve
3	Maximum power of 24-hour load curve (kW)	10	Kurtosis of 24-hour load curve
4	Minimum power of 24-hour load curve (kW)	11	Mode of the five-bin histogram for a 24-hour load curve (kW)
5	Range of power, i.e., maximum – minimum (kW)	12	Longest period of successive increase (h)
6	Percentage fraction of values above mean	13	Longest period of successive increase above mean value (h)
7	Sum of net-loads during business hours, i.e., 09:00–17:00 (kW)	14	Number of peaks

These features serve to represent principal characteristics of an electricity load curve into distinct feature values. The details of these features are explained below and justified in [24]:

- *No. 1 – 5:* The first five features are basic numerical measures commonly used in quantitative data analysis [26].
- *No. 6:* The fraction of values above mean refers to the percentage of values in the load curve which is larger than the average value. This percentage is input in fractional form and thus, has a range of 0 to 1.
- *No. 7 – 8:* The sum of net-loads during business hours (feature No. 7) and non-business hours (feature No. 8) are total sums of all data points within the specified time periods, and represent the amount of feeder-level power consumption/generation within and outside the period from 09:00 to 17:00, respectively.
- *No. 9 – 10:* Skewness and kurtosis measure the shape of the data point distribution, in comparison to a normal distribution curve. The kth standardised moment denotes the ratio of kth moment about the mean to the kth power of the standard deviation. Skewness and kurtosis will have $k = 3$ and $k = 4$ respectively. Skewness measures the asymmetry of the distribution about its mean value, for which a positive skewness value in a 24-hour load curve means that the data is right-tailed, and the mean value is larger than the median value, and vice versa. Kurtosis measures the sharpness of the peak of data distribution so as to indicate the prevalence of extreme values. A positive kurtosis value, i.e., heavy-tailed distribution, represents a sharper peak and fewer extreme data points, while a negative kurtosis value represents a more rounded peak and the presence of extreme data points. Both of these measures represent additional information about the shape of the input load profile.
- *No. 11:* The mode of five-bin histogram for a 24-hour load curve is obtained by firstly dividing the range of data point values

[3] The timeframes for each season are based on the Met Office’s definitions of meteorological months [23] summer are June–August, winter are December–February and transitional seasons of spring and autumn are Mar–May and Sep–Nov respectively.

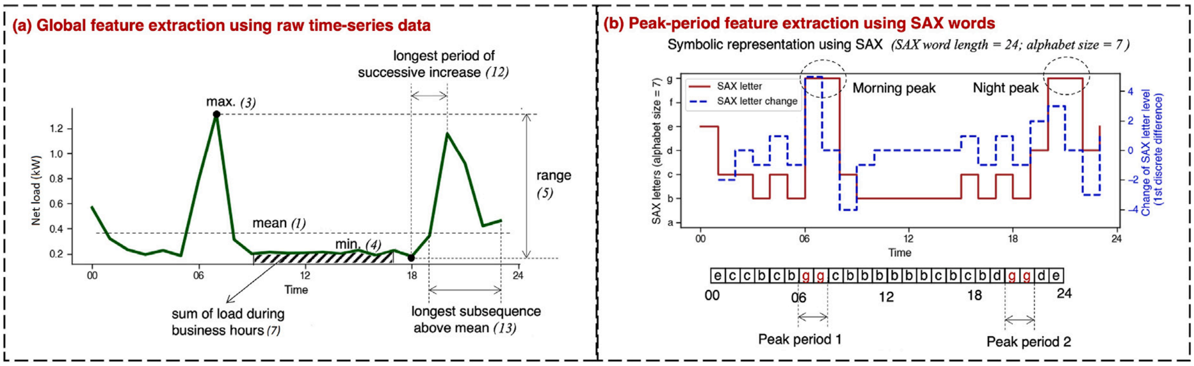


Fig. 2. Figures depicting how specific IFEEL input features (numbered based on Table 2) describe characteristics of an example net load curve [24].

into five bins of the equal size, and then counting the number of data points in each bin. The average value of the bin with the most data points is defined as the mode which will be taken as the input feature. For a 24-hour load curve, it is likely that every data value is unique, and thus, also the mode value. By counting the frequency of value ranges, rather than individual values, feature no. 11 is more useful than the conventional mode metric in determining the location of load curve's mode.

- *No. 12–13:* The longest period of successive increase and longest period of successive increase above mean value are self-explanatory, in which the latter requires all data point values in the sequence to be above the mean value. These features describe the temporal lengths of load dips and peaks. A large value of feature no. 12 could be represent either a dip or a peak occurring over a longer duration, while a large value of feature no. 13 would represent a peak occurring over a longer duration.
- *No. 14:* The number of peaks is annotated by the symbolic aggregate approximation (SAX) representation technique [27]. The SAX converts a time-series dataset into a string of letters and each letter represents the distance of a data point from the mean. For the detailed procedures of using the SAX representation technique to represent the number of peaks, refer to Appendix and [24,27].

As mentioned in Section 2.1, two additional features were input into the DPVSCE model: the intra-week group and the season (summer, winter or transition seasons) of data collection. To visualise the consumption patterns these features represent, 24-hour net-load curves averaged from 81 households with various PV penetration rates are presented in Fig. 3. Details of the dataset used are explained in Section 3.3 below. From Fig. 3a, the typical household electricity consumption pattern can be observed: a dip after midnight, followed by peaks in the morning and evening. With increasing PV penetration rates, it can be seen from Fig. 3 that midday PV generation increases, causing larger dips during daylight hours. Fig. 3 also presents distinct household load consumption patterns for each season and intra-week group. Specifically, the following four common observations can be seen.

(1) With no DPVS installed, the electricity demand generally increased from summer to winter through the 'transitional' season, due to the increased usage of heating appliances. *(2)* For the intra-week groups, a common pattern emerged during the working hours from 09:00 to 17:00 when electricity consumption decreased in the order of Sundays, Friday–Saturdays and Monday–Thursdays. This is likely due to residents leaving their homes for work during weekdays and staying at home during weekends. Note that the data provided were pre-COVID, and thus did not reflect rising levels of remote working. *(3)* The sharpest gradients during the morning and evening peaks were observed in winter, followed by the transitional seasons and summer. The morning and evening peaks arose when residents wake up and return home respectively. Such peaks were especially pronounced in colder and darker seasons due to increased usage of appliances, such as kettles and lighting. *(4)* the difference in evening peaks between intra-week groups in each season was caused by the reduced evening peaks of 'out nights', leading to less sharp evening peaks on Fridays and Saturdays. Besides, evening peaks appeared later in the group of Monday–Thursday as evening household electricity consumption only began when residents return home from work.

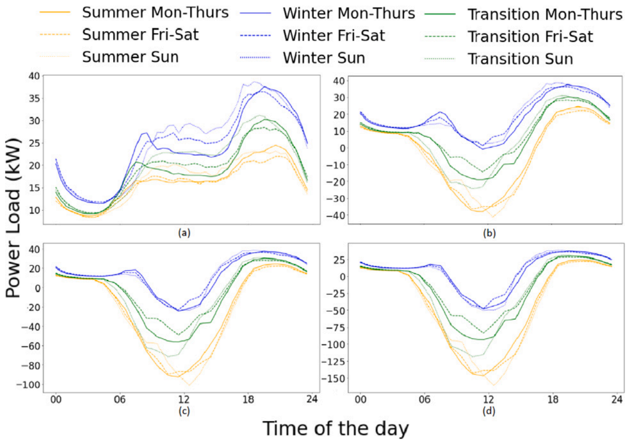


Fig. 3. 24-hour net-load curves averaged from 81 household consumption data with PV penetration rates at (a) 0%, (b) 33%, (c) 67% and (d) 100%. The yellow, blue, and green lines indicate the summer, winter, and transitional seasons, respectively. The solid, dashed, and dotted lines indicate Monday–Thursday, Friday–Saturday, and Sunday, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Therefore, given the input features of the season and intra-week groups, the ANN will 'learn' the shape of typical household load consumption of the specific season and intra-week group, in addition to their IFEEL features. For instance, the range of power load (IFEEL feature No. 5) would be lower for Friday and Saturday, compared to other intra-week groups within the same season due to the smaller evening peak.

2.3. Artificial neural networks

For the proposed method, an ANN was used to map the 16 extracted features to the estimated DPVS capacity. An ANN is an ML algorithm which aims to learn a non-linear function from a set of input values to output values as

$$f(\cdot) : X^m \to Y, \forall m \in M, \tag{4}$$

where $f(\cdot)$ is the non-linear function parameterised by the ANN, X^m is the set of input samples of the feature $m \in M$, M is the feature set, and Y is the set of output values, i.e., training labels. Each set of the input feature m contains I samples and each sample is denoted as X_i^m. These samples make up the training dataset, which can be described as

$$D : \left\{X_i^m, Y\right\}, \forall i \in I, m \in M, \tag{5}$$

where D is the training dataset. With known training labels, the ANN serves as a supervised learning algorithm to predict the training labels.

The basic structure of an ANN consists of three main components: the input layer, hidden layer(s), and output layer. The input layer contains a set of neurons, each representing an input feature value. Each hidden layer contains a set of neurons which converts the values in the previous layer by the following function

$$f_n^{\text{out}} = \sigma(\sum_{n=1}^{N} w_n \cdot f_n^{\text{in}} + b_n) \tag{6}$$

where f_n^{out} and f_n^{in} are the output and input of the neuron n in each hidden layer respectively, σ is the non-linear activation function, w_n is the weight of the neuron n in each hidden layer and b_n is the bias term. The output layer has a single neuron which also performs the operation shown in Eq. (6), without the activation function. In an ANN, the flow of information from the input layer to the output layer is defined as forward propagation. The weight values of an ANN are determined by backpropagation [28], through which a loss function, which measures the error between the predicted value and the training label, is minimised by propagating backwards to update weight parameters over defined iterations.

The advantages of using an ANN in the proposed research can be summarised as *(1)* ANNs have the ability to learn complex non-linear models and represent functions with any form of shape [29]; *(2)* ANNs are able to learn in real-time with updated data; *(3)* ANNs are less susceptible to the issue of local minimums compared to other ML algorithms, due to the low probability of having all input features being at their optimal values at a single point in the cost function space [30]. On the contrary, the disadvantages of ANNs are that *(1)* the randomness of weight initialisation would lead to inconsistent prediction performances; *(2)* ANNs are sensitive to the scaling of input feature values; *(3)* ANNs require manual tuning of hyper-parameters. These disadvantages will be addressed in the simulation methods, as explained in the following section.

3. Simulation methods

This section explains the methods, dataset and evaluation metrics used for ANN hyper-parameter optimisation and model sensitivity analysis.

3.1. Hyper-parameter optimisation

The hyper-parameter optimisation process aims to tune the ANN towards the best prediction performance. The activation function, number of hidden layers, and the optimiser[4] were selected based on qualitative reasoning, while the number of hidden neurons, initial learning rate,[5] L2 regularisation value,[6] and maximum iterations were tuned by quantitative analysis.

The ReLU function was selected as the activation function due to its preservation of linear properties which generalise well and simplifies gradient-based optimisation [31]. Next, the number of hidden layers was fixed at two, since two hidden layers allows the model to represent a function of any shape, and there is no theoretical reason to use any more hidden layers [29]. In fact, increasing the number of hidden layers increases the risk of undesirable over-fitting [31].

The Adam solver [32] was used to optimise the weights of the ANN as it has little memory requirement for first-order gradients and is compatible with non-stationary and sparse gradients. The key feature of the Adam solver is that it uses the exponential moving average of the gradient and the squared gradient, instead of the gradient itself, to update the weight parameters. This quickens convergence and smooths the gradient descent process [32]. In addition, at each iteration, the Adam solver calculates the gradient-related values using random subsamples of the provided dataset, instead of the entire dataset, thus further speeding up the computation process.

For quantitatively optimised hyper-parameters, the number of neurons in each hidden layer was optimised by finding a trade-off between the computational time and prediction accuracy. Increasing the number of hidden neurons enables a higher representational capacity of the ANN, potentially improving prediction accuracy, but also increases the computational time exponentially. Initial learning rate and L2 regularisation value were optimised using the exhaustive grid search technique [33], until the same set of values appeared twice in a row. The number of maximum iterations was optimised by observing root mean squared error (RMSE) values from both training and validation sets, and a divergence of the two values would indicate the model overfitting to the training set [34], and thus poor generalisation performance.

3.2. Sensitivity analysis

This subsection explains the methods used to analyse the proposed DPVSCE model's sensitivity to three variables: *(1)* the temporal characteristic of data collection, *(2)* the number of households served by a substation, and *(3)* the proportion of PV-equipped properties. The sensitivity to these variables was also examined under various PV penetration rates. Analysis of the first variable aimed to decipher the effects of using data from specific seasons or intra-week groups, and it was also crucial to understand how the model's performance would change with respect to the other two variables, since they are common differences between different substation feeders.

3.2.1. Temporal characteristic of data collection

To test the model's sensitivity to the temporal characteristic of data collection, the estimation accuracy was compared between only using data samples of each season or each intra-week group.

For the season-based analysis, the base dataset was filtered based on their season. For example, for summer, only the data samples from summer (i.e., season input feature value = 1) were used. After obtaining the filtered data samples, the filtered training dataset was standardised and the scaling (mean and standard deviation of each feature) was applied to the testing dataset. This was to ensure that there was no information leak from the testing dataset to the training dataset, which might have resulted in the ANN being biased towards the testing dataset, and thus produce unrepresentative results. Using these standardised datasets, the ANN, with the optimal hyper-parameters obtained earlier, was trained and tested for its prediction performance.

Similarly, for intra-week groups, the process was analogous to the season-based analysis above, except instead of filtering the base dataset based on seasons, samples from specific intra-week groups were used. This intra-week group analysis served to test the effectiveness of the proposed intra-week groups in extracting knowledge on household load consumption behaviours.

3.2.2. PV penetration rate

The average and standard deviation of prediction RMSE at each true DPVSC value were recorded and plotted for each season and intra-week group. This created a residual plot which depicts the variation of prediction performance in response to varying PV penetration rates. This tested the model's prediction performance for substation feeders which serve residential neighbourhoods with different amounts of DPVS installed.

[4] This hyper-parameter determines how weight parameters are updated at each iteration.

[5] This hyper-parameter determines how weight parameters are updated at each iteration.

[6] This hyper-parameter prevents model overfitting by suppressing the size of weight parameters.

3.2.3. Number of households

The present research also assessed the model's effectiveness for substation feeders which serve a different number of households. The pseudocode for the sensitivity analysis to the number of households is shown in Algorithm 1. In this research, the number of household properties tested was from 20 to 80, in multiples of 10, and the tested number of DPVS installed was from zero to the total number of household properties. This process was also repeated using only data samples of specific seasons and of specific intra-week groups respectively. Ultimately, this provided a comprehensive overview of the proposed algorithm's performance in terms of four variables: PV penetration rate, number of household properties, the use of only season-specific data samples, and the use of only intra-week group-specific data samples.

Algorithm 1 Sensitivity analysis to the number of household properties

```
input: 81 individual household load consumption datasets (annual),
  PV generation dataset (annual)
  for no. of household properties (20 to 80, in multiples of 10) do
    for no. of DPVS installed (0 to no. of household properties): do
      Randomly select the specified number of individual household
      load consumption datasets
      Add PV generation of a specific number of installed DPVS to
      load consumption
      Obtain annual feeder-level net-load dataset
      Extract averaged 24-hour net-load datasets
      Extract IFEEL input features
    end for
    From for loop above, obtain a base dataset of samples
    Split dataset into training and testing datasets
    Select data samples based on season or intra-week group (if
    necessary)
    Standardise training and testing datasets
    Train ANN using training dataset
    Test ANN using testing dataset and record RMSE of estimations
  end for
output: $m \times n$ array of RMSE, where $m$ is the number of households and
  $n$ is PV penetration rate
```

3.3. Simulation setup

The proposed simulation required a dataset which was derived from two components, i.e., household load consumption data and PV generation data. The hourly household load consumption data were collected from 162 individual households based in central London, for the entire year of 2013. The PV generation data were obtained from the Photovoltaic Geographical Information System (PVGIS) [35] with manually-defined parameters shown in Table 3. The SARAH solar radiation database was used to generate the PV generation data as it was the only database on PVGIS which covered the UK. The azimuth angles chosen reflect south-facing PV panels, since these produce higher PV output in the Northern Hemisphere. Since a single substation feeder typically serves a maximum of 75 households [36], the maximum number of households per feeder was set as 81, so that the 162-household load data could be used to represent two separate sets of the substation feeder load. With the average residential PV capacity at 4.2 kWp [37], the total capacity of 81 residential PV systems would be 340.2 kWp, which was set as the peak capacity for the PV generation dataset. The published ML-ready dataset, as mentioned in Section 1, thus contains samples representing 81-household substations with varying PV penetration rate.

The proposed model was written largely using the *scikit-learn* ML package [38]. The data were split as 70% for the training set, 20% for the validation set, and 10% for the testing set.

Table 3
Manually defined parameters for producing PV generation dataset.

Location	Year	Azimuth	Peak Power
City of Westminster	2013	45°/0°/−45°	340.2 kWp
PV Technology	**Mounting**	**Roof Slope**	**Database**
Crystalline Silicon	Fixed	30°	PVGIS-SARAH

The RMSE was used as the metric of the learning loss as

$$l = \sqrt{\frac{1}{I}\sum_{i=1}^{I}\left(y_i - \hat{y}_i\right)^2}, \tag{7}$$

Considering possible inconsistencies and randomness, the simulation was repeated five times and the average performance was measured using the three metrics. In addition to the RMSE, two other performance metrics were used: mean absolute error (MAE), which is represented by Eq. (8), and standard deviation of absolute error (AE).

$$\frac{1}{N}\sum_{n=1}^{N}\left|y_i - \hat{y}_i\right|. \tag{8}$$

The MAE served as an additional performance metric, where high error values are not penalised as much as with the RMSE. The standard deviation of AE measures the spread of the predictions' absolute error. Both metrics reflect the preciseness of predictions: a large difference between MAE and RMSE indicates the presence of substantial error values, while a high standard deviation of AE indicates a large range of absolute error values. Note that an inaccurate (high MAE) but precise (low standard deviation of AE) algorithm would still be useful for DPVSCE as the true PV capacity value can be reliably inferred from the estimated value via the known consistent error value.

4. Simulation results

In this section, the results of performed simulations are presented. The results of the hyper-parameter optimisation process are first presented to explain the rationale behind each optimal hyper-parameter value. Then, the results of sensitivity analysis are presented and their implications for real-world application of the proposed DPVSCE method are discussed.

4.1. Evaluation of hyper-parameters

The evaluation result for the number of neurons in each hidden layer is presented in Fig. 4 which illustrates the average training time and RMSE against the number of neurons in each hidden layer. The standard deviation of training time and RMSE are also plotted as a range around the average values. It can be seen that the standard deviation of both training time and RMSE are within a small range for each number of hidden neurons. The learning accuracy increases with the cost of increased training time as the number of hidden neurons increases. In particular, when the number of hidden neurons exceeded 300 per hidden layer, the training time increased drastically with minimal improvement in prediction accuracy. Hence, the number of hidden neurons is selected at a point where training time remained at 150 s before rising sharply and the RMSE value began to flatten at 26 kW, i.e., 3000 neurons per hidden layer.

The evaluation results for optimising the initial learning rate and L2 regularisation value are presented in Table 4, where the tested hyper-parameter values are listed and the best performing values are highlighted in the bold font. Note that the initial learning rate has already been optimised by the second iteration since the best-performing value had remained constant for two iterations. Despite this, a third iteration was performed as the L2 regularisation value was still being optimised. This additional iteration further corroborated the optimum

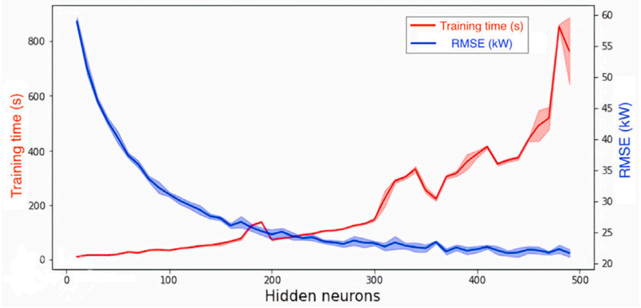


Fig. 4. Average training time and RMSE against the number of hidden neurons. The blue line is the average RMSE and the red line is the training time. The standard deviation is depicted as a range around the average values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Evaluation of the initial learning rate and L2 regularisation value with the progression of iterative optimisation, with the best performing value in **bold**.

Iteration	Initial learning rate	L2 regularisation (in 10^{-5})
1	**0.01**, 0.001, 0.0001	100, 10, **1**
2	0.1, **0.01**, 0.005	**5**, 1, 0.1
3	0.05, **0.01**, 0.075	7.5, **5**, 2.5

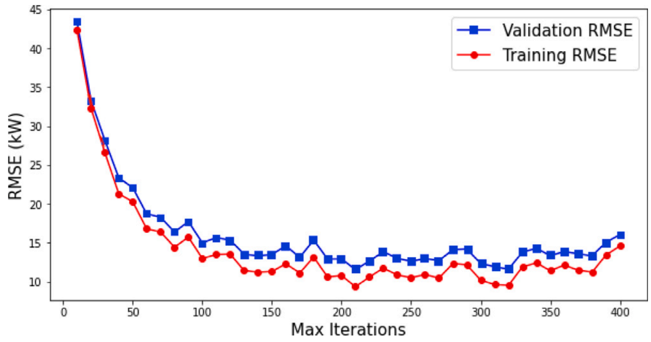


Fig. 5. Training and validation RMSE against maximum training iterations. The blue line is the validation RMSE and the red line is the training RMSE. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of 0.01 for the initial learning rate and thus, was selected with 5 × 10^{-5} as the optimal L2 regularisation value.

The final hyper-parameter to be optimised was the maximum number of iterations during the ANN training process. Fig. 5 shows the average training and validation RMSE against the maximum number of iterations. As expected, the ANN model performed better on the training samples than the validation samples, given that it was trained using the latter, thus, it has learned the relationship between the input features and the output variable. Prediction accuracy was also observed to decrease in a logarithmic fashion as maximum iterations increased, although it stagnated at an RMSE value of roughly 15 kW at more than 210 max. iterations. This result makes any more iterations unnecessary.

To make a more informed decision on the maximum number of iterations, the training duration was taken into consideration. For the assessment of training duration, the mean and standard deviation of the training time was plotted against the maximum number of iterations in Fig. 6. The aim was to minimise both RMSE and training time, with priority placed on minimising the former. It can be observed from Fig. 6 that the training time stayed within the range of 75 to 100 s when the maximum number of iterations was between 110 and 400. Thus, based on the minimum validation RMSE value, the maximum iterations was retained at 210.

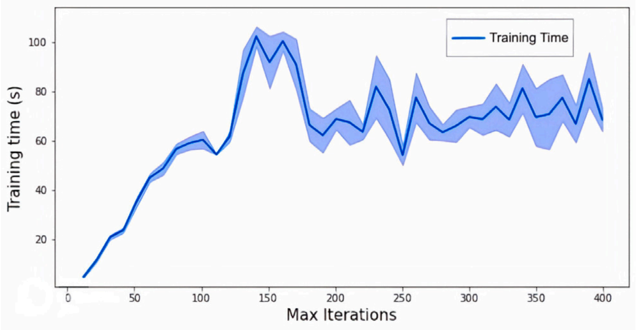


Fig. 6. Average training time against maximum training iterations. The blue line is the average training time, with the standard deviation plotted as a range around the average values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Finalised hyper-parameter used on artificial neural networks for the sensitivity analysis.

Hyper-parameter	Value
Activation Function	ReLU
Back-propagation Method	Adam Solver
Number of Hidden Layers	2
Number of Hidden Neurons (in Each Layer)	300
Initial Learning Rate	0.01
L2 Regularisation Parameter (Alpha)	5 × 10^{-5}
Maximum Iterations	210

Table 6
Estimation performances by prediction models using only data samples from specific seasons.

	RMSE (kW)	MAE (kW)	Std. of AE (kW)
Summer	9.83	7.43	6.42
Winter	18.40	13.70	12.30
Transitional	13.10	8.33	10.10
All Seasons	12.00	7.87	9.08

Therefore, the optimal ANN hyper-parameters are summarised in Table 5. These custom hyper-parameter values were used for the subsequent experiments, with all other unmentioned hyper-parameters fixed at the default values provided by the *scikit-learn* package [32].

4.2. Results of sensitivity analysis

4.2.1. Sensitivity to seasons

For the sensitivity analysis to seasons, only data samples from specific seasons (i.e., summer, winter and the transitional seasons) were used in training each prediction model. The average RMSE, MAE, and standard deviation of AE for each circumstance are shown in Table 6.

Two key observations can be made from Table 6: *(1)* Using only data samples based in the winter months produced the worst results across all three metrics, while the other three seasonal groups' performances were relatively close to one another. *(2)* Comparing between summer, transitional seasons, and all seasons, the RMSE and standard deviation of AE increased, while the MAE decreased in the order of the transitional seasons, all seasons, and summer. Using only data samples from summer produced the most accurate predictions, with a cost of reduced preciseness and vice versa for the transitional seasons whereas using data samples across all four seasons achieved a balance between the two.

The first observation can be explained by the low amounts of PV generation during winter, resulting in less obvious distinctions between the net-load curves of different installed PV capacities. This is reflected in Fig. 3, where the net-load behaviours based in winter changed the least with increasing PV penetration rate. These characteristics

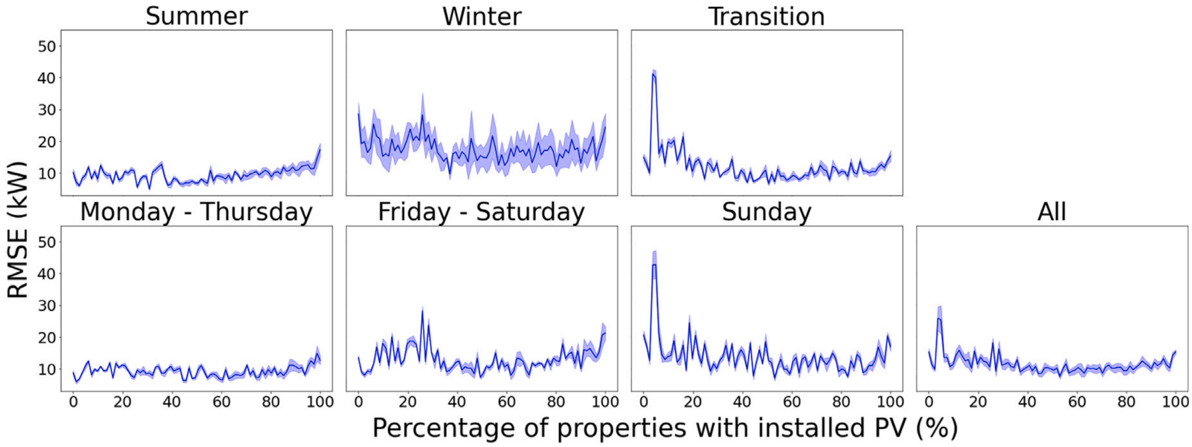


Fig. 7. Average RMSE of estimations against PV penetration rate using all data samples and data samples from specific seasons and intra-week groups, with standard deviation depicted as a range around the average values.

Table 7
Estimation performances based on different metrics using only data samples from specific intra-week groups.

Intra-week group	RMSE (kW)	MAE (kW)	Std. of AE (kW)
Mon–Thurs (A)	9.50	7.23	6.16
Fri–Sat (B)	13.80	9.67	9.79
Sun (C)	15.20	9.51	11.90
All groups	13.00	8.90	9.51

imply a poorer prediction performance during winter. Thus, DNOs are strongly encouraged to have a robust winter data repository to ensure an accurate and precise ANN for winter predictions.

It is important to note that there were roughly twice as many data samples from transitional seasons (6 months of data) as that from either summer or winter (3 months of data) individually. This difference in the amount of training data could have allowed the ANN to better predict the relationship between the input and output variables. To test this hypothesis, the experiment process was repeated using a half of the training and validation data samples (selected randomly) from transitional seasons and the performance results were 17.50 kW, 10.80 kW, and 11.90 kW for RMSE, MAE, and standard deviation of AE respectively, thus confirming the aforementioned hypothesis.

With equal number of training and testing samples, the prediction performance of a transitional-season model is now closer to that of a winter model, supporting the theory that the difference in PV generation between seasons causes the difference in the estimation performance. This also suggests that increasing the number of data samples used for training can greatly improve the prediction performance, which will be especially useful for predictions in winter. As such, DNOs may wish to expand the storage and maintenance capabilities of a large training data repository or collaborate with each other to share relevant feeder level net-load data. As mentioned earlier, more focus can be placed on winter data storage as winter-based prediction performance lag behind summer. In fact, the results showed that using only summer-based data produces comparable performances to using the data from all seasons, despite requiring only a third of data quantity.

4.2.2. *Sensitivity to intra-week groups*

Further analysis only used data samples from specific intra-week groups. The average RMSE, MAE, and standard deviation of AE of each circumstance are shown in Table 7.

For simplicity, the intra-week groups will be referred to by the bracketed letters shown in Table 7 (“A” – “C”). Two key observations can be made from the results: *(1)* Group A produced the best results across all 3 metrics, compared to using all samples, as presented in both Tables 6 and 7, *(2)* Comparing between the three intra-week groups, performance metric values generally increased, reflecting a reduction in both accuracy and precision, in the order of Groups A, B and C. The first observation leads to the straightforward implication that the priority should be placed on utilising feeder-level net-load data from Monday to Thursday for model tuning and predictions. Also, to further improve prediction performance on other intra-week groups, storage and maintenance of a larger data repository for Groups B and C may be required for model training.

There are two hypotheses which could explain the second observation. The most straightforward explanation would be that the household consumption behaviour was the most consistent during the days of Group A, followed by Groups B and C, which allowed the differences in net-load curve to be solely based on differences in DPVSC and PV generation. The second hypothesis is that the amount of data which constitutes each data sample affects prediction performance. Specifically, IFEEL features obtained from the Group A net-load curve were the average of four days’ worth of data, while the Group B net-load curve was the average of only two days’ worth of data and just one day for the Group C. Considering this factor, the results could reflect the effectiveness of averaging 24-hour net-load curves to reduce the daily variability of weather conditions.

To further test this hypothesis, more in-depth research could be conducted by maintaining the intra-week group variable constant and comparing prediction performance when averaged net-load curves are obtained from different quantities of data. For example, three IFEEL datasets could be created using IFEEL features extracted from the Group A averaged net-load curves of four, eight, and 12 days respectively. Prediction performance would then be assessed for these three circumstances. This analysis should provide more conclusive results on the effectiveness of removing noise in the data by taking the average of more 24-hour net-load datasets.

4.2.3. *Sensitivity to PV penetration rate*

The relationship between PV penetration rate and prediction error, when using only data samples of specific seasonal or intra-week groups, is shown in Fig. 7. Three key observations can be obtained: *(1)* RMSE values for all plots tend to be higher and more inconsistent at PV penetration rates below 50%, compared to above 50%. *(2)* Large increases in RMSE values tend to appear at 100% PV penetration rates and *(3)* the results presented in Tables 6 and 7 are consistent with this plot, where a lower curve on the y-axis implies more accurate predictions and the presence of sharp spikes and dips implies imprecise predictions. For example, comparing between summer and the transitional seasons plots in Fig. 7, the RMSE values for summer were generally lower than that for transitional seasons, but contained more extreme values, representing its higher accuracy but lower precision.

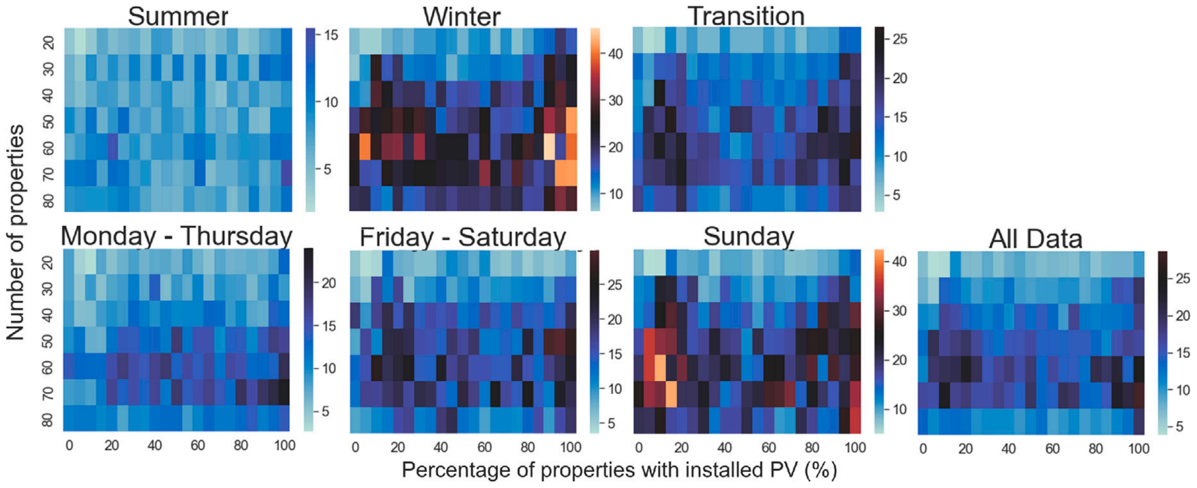


Fig. 8. Average RMSE of predictions varying with the number of properties and percentage of properties with installed PV (based on the true value of installed PV capacity), with the title of each heatmap indicating the season or intra-week group of the data samples used.

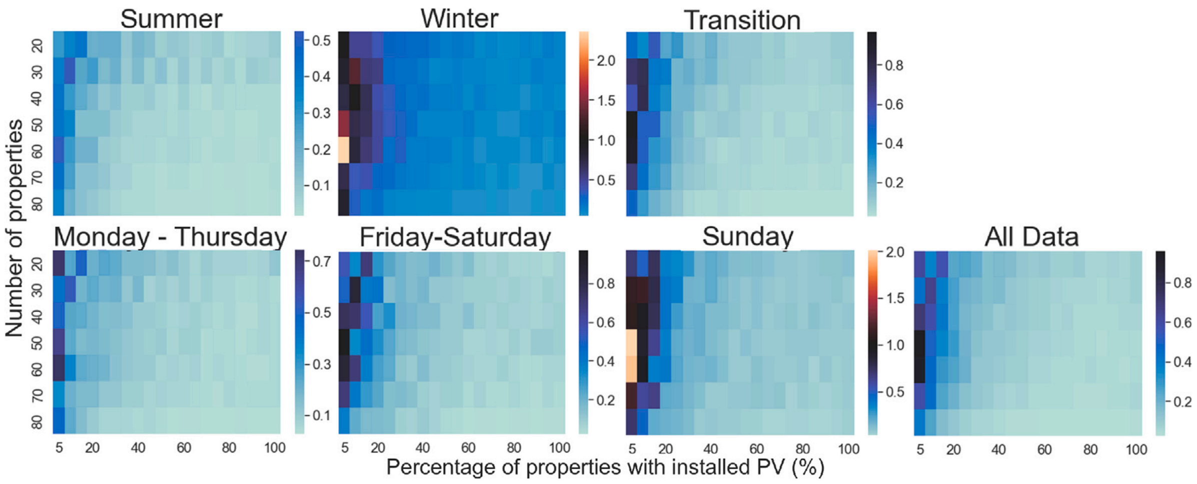


Fig. 9. Average mean absolute percentage error of predictions varying with number of properties and percentage of properties with installed PV, with the title of each heatmap indicating the season or intra-week group of the data samples used.

The first observation could be explained with the low PV generation at low PV penetration rates. With lower PV generation, it was easier for inconsistent weather patterns or load consumption behaviours to be mistaken by the algorithm as changes in installed DPVS capacity. For example, the holiday season could lead to a significant spike in household load consumption which appears similar to a reduction in DPVSC at low PV capacities, leading the algorithm to underestimate the PV capacity. At high PV capacities, however, such a consumption spike might be minimal compared to the amount of PV generated, and thus would not be detected by the DPVSCE algorithm.

With this result, DNOs may wish to prioritise employing the proposed DPVSCE method on newer estates with higher PV penetration rates for more accurate predictions, given that only 4.1% of UK households having installed solar panels in 2023 [39]. Otherwise, obtaining an initial estimate of the PV penetration rate may be useful in determining the expected error range to decide if the proposed method's estimation can be utilised. This initial estimate could be performed either using satellite imagery or in-person surveying, though the latter will be significantly more resource-intensive.

From the second observation, the next step is to understand if the highly erroneous predictions are specific to a true PV capacity of 340.2 kW (see Section 3.3) or if they are specific to the condition of a 100% PV penetration rate. This is explored in the next section where the prediction performance of a varying number of households was measured. The number of households results in different PV capacities at the 100% PV penetration rate.

4.2.4. Sensitivity to the number of households

The effects of number of households and PV penetration rate on prediction error are presented in Fig. 8, with each heatmap representing only data samples from specific seasons or intra-week groups. Note that the colour scale of each heatmap is constant, with black corresponding to a value of 25 kW, warmer colours being more than 25 kW and cooler colours being less than 25 kW. Each heatmap also has a colour scale which indicates the range of values present in the heatmap. Three key features can be observed from Fig. 8, which are common patterns for all sub-figures: *(1)* The phenomenon of highly erroneous predictions at near 100% PV penetration is observed across different numbers of households. Combined with high RMSE values generally seen at near 0% PV penetration, this can be visualised as a quadratic curve when plotting RMSE against PV penetration rate (see Fig. 7); *(2)* There is a general trend of increasing RMSE values as the number of households increases; *(3)* Despite the trend mentioned previously, the RMSE drops sharply between 70 and 80 households served by the substation feeder.

The first observation can be interpreted as a non-linear residual plot, meaning that the relationship between predicted installed PV capacity and PV penetration rate has not been properly detected by the model.

The solution usually entails either transforming an existing input feature or adding a "missing" feature [40]. A possible input feature would be the number of households or the theoretical maximum DPVSC based on the number of households. Regardless, a high PV penetration rate at more than 80% is unrealistic as the costs of integrating that much DPVS into the grid are predicted to suppress the PV penetration to roughly 60% [41]. Thus, this observed limitation of the proposed model will unlikely become a concern in real-life applications. The second observation could be explained by a possible cause–effect relationship between the true PV capacity and margin of error, where an increase in the former leads to an increase in the latter. To test this hypothesis, the mean absolute percentage error (MAPE) [42] is plotted as a heatmap in Fig. 9. The colour scale is constant where black represents the value of 1.0, with warmer colours representing values more than 1.0 and cooler colours representing values less than 1.0.

It can be seen that the error-to-true value ratio stayed relatively constant for all numbers of households at the PV penetration rate of 40% and above, while at below 40% PV penetration, MAPE tends to increase with a decreasing number of households and decreasing PV penetration rate. With homogeneous MAPE values at about 0.1 when using data from specific seasons or intra-week groups (see Fig. 9) at higher PV penetration rates, this provides a useful guide for expected error margins in predictions. Based on this guide and acceptable error margins, DNOs can then decide on the suitable substations to use this method on, depending on the substations' estimated number of households and PV penetration rates. However, high MAPE values at lower PV penetration rates mean DNOs cannot solely rely on the proposed method for DPVSCE. Instead, the proposed method can serve as an additional estimation to other DPVSCE methods, such as satellite imagery-based methods or manual checking of installed solar panels.

Finally, the third observation is likely to be the result of overfitting the hyper-parameters to the specific circumstance of 80 households, causing the algorithm to perform poorly when the number of households is different. To tackle this, DNOs can employ the proposed method by either training separate predictors for feeders serving a varying number of households, or including data samples corresponding to various numbers of households in the training process, to create a predictor which is able to generalise well.

5. Discussion

In this paper, extensive analysis was conducted to assess the proposed method's sensitivity to changing circumstances, this section summarises the few key takeaways to note.

Firstly, the amount of data will be crucial to improving prediction performance, with more training data samples leading to more accurate and precise estimations. In this case, doubling the number of training data samples (averaged intra-week net-load curve) from about 6396 to 12,792 saw a reduction in average RMSE by 25.1% from 17.5 kW to 13.1 kW for the model trained using only transitional season data. To put this into perspective, this error reduction is larger than the average residential PV system capacity of 4.2 kWp [37]. DNOs can use this insight to perform relevant cost benefit analysis to determine if the error reduction justifies the costs of storing additional data.

Secondly, this paper has shown that there is merit in developing separate prediction models for specific time periods. For example, despite only using about 6396 training samples for the summer-based model, compared to 25,584 samples for the model trained with all data, the former performed comparably well with the latter, with RMSE values of 9.83 kW and 12.0 kW respectively, and MAE values of 7.43 kW and 7.87 kW respectively. Such a model would have the dual advantage of being accurate and requiring little data. Similarly, in terms of models trained and tested exclusively using specific intra-week groups, using data from Mondays to Thursdays performed the best with an average RMSE of 9.50 kW, followed by Fridays to Saturdays with 13.8 kW and Sundays with 15.2 kW. These results could be caused by either consumers exhibiting the most predictable consumption patterns from Mondays to Thursdays, the effectiveness of taking the average daily net-load curve of a longer time-series, or both. The exact contributing factor could be determined through further research.

Finally, further efforts in performance improvements should focus on making estimations for substations serving fewer households, where higher percentage errors were observed. Using the summer-based model as an example, MAPE values only stabilised at roughly 0.1 at 60% PV penetration for 20 households, while the MAPE value for 70 households dropped much more rapidly as PV penetration rate increased. Thus, DNOs should factor the number of households into account when performing ML-based DPVSCE, noting higher error margins at fewer number of households. Regardless, this result serves as an useful error margin guide for DNOs to make more informed decisions on the utility of the proposed DPVSCE method based on the estimated number of households and PV penetration rate served by the substation feeder.

5.1. Limitations

In conjunction with the experiment results, it is crucial for stakeholders to be aware of the limitations of this paper's methodology.

Firstly, in the proposed method, each data sample only contains data from within one week, future research can be conducted to assess the effectiveness of removing net-load curve noise by taking the average of data across multiple weeks. If using more data to obtain averaged net-load curves is indeed more effective, note that this will incur the cost of reduced training samples, given that there is a constant amount of time-series data available. This cost might, in turn, lead to worsened prediction performance. Thus, this leads to a second potential research area of exploring the trade-off between the number of training samples and the amount of data used per sample.

Next, the rising popularity of residential energy storage systems [43] will lead to significant changes in net-load curve behaviour due to the load shifting effects of the battery charging during peak PV production and battery discharging during peak electricity consumption [44]. Conversely, with huge growth in the electric vehicle (EV) market [45], this will further increase evening peak load when residents return home to charge their EVs [46]. Thus, there will be a need for further research in updating the proposed DPVSCE method to incorporate the effects of behind-the-meter energy storage and EV charging when analysing feeder net-load curves.

Also, the data used in this research was from before the COVID-19 pandemic, and the lockdown and subsequent transition to remote working have changed residential electricity consumption patterns, most notably a decrease in overall demand and a reduction in morning and evening peaks [47–49]. Tuning the model to changes particular during the COVID-19 pandemic will be a crucial area for future explorations.

Finally, the sensitive nature of load consumption data meant that commercial-related limitations had rendered it difficult to source for ground truth data which, in this case, is real substation load data with known amount of installed DPVS capacity. As such, this paper does not contain assessment of the proposed method based solely on real life data. DNOs will need to assess the trained model with their own ground truth data.

6. Conclusion

With behind-the-meter DPVS in the UK having more than doubled over the last decade, this calls for an increasing need for updated information on installed DPVS capacity to ensure the continued smooth operations of the power networks. However, currently researched DPVSCE methods face a major limitation of having extensive data requirements, whether they are satellite images, household-level net-load datasets, or weather data. The proposed method developed in this paper overcomes this limitation by utilising a newly proposed intra-week demand grouping and employing the data pre-processing technique of extracting time-invariant IFEEL [24] features from averaged 24-hour net-load curves. This model is highly accessible for DNOs or other relevant stakeholders since it only requires feeder net-load consumption data (with no specific time resolution requirement) as input to perform DPVSCE. The proposed model also allows DNOs to be self-sufficient in terms of the required data to perform DPVSCE, by negating the need for third-party data such as satellite imagery or weather data.

In addition to making available an ML-ready dataset representing substations with varying PV penetration rate, this paper also contributes by conducting a holistic assessment of the deep neural network as a tool for estimating installed DPVS capacity linked to a substation. This assessment involves testing model estimation accuracy and precision, based on 3 different metrics: root mean squared error, mean absolute error and mean absolute percentage error, in response to varying data collection time periods, number of households linked to a substation and percentage of households with installed PV (PV penetration rate). The sensitivity analysis revealed that operating the model exclusively in summer can produce more accurate estimations with less training data, providing an 18.1% improvement in accuracy while using only a-third of the data volume compared to an all-year model. Similarly, exclusive model operation from Mondays to Thursdays reduced estimation error by 26.9%. Ultimately, improved DPVSCE capabilities will enable cost savings related to data acquisition and storage, whilst providing crucial insights for DNOs to maintain safe and efficient operations of electricity distribution systems in the midst of a changing electricity generation landscape.

CRediT authorship contribution statement

Lingxi Tang: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation. **Masaō Ashtine:** Supervision. **Weiqi Hua:** Supervision. **David C.H. Wallom:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Appendix. Symbolic aggregate approximation

The SAX consists of three procedures: *(1)* Z-score normalisation, *(2)* piecewise aggregate approximation, and *(3)* discrete representation, which is explained as follows.

First, the time-series dataset is standardised using Z-score normalisation as

$$z_t = \frac{x_t - \bar{x}}{\sigma}, \tag{A.1}$$

where z_t is the normalised data point at the time step t, x_t is the original data point at the time step t, $\bar{x}$ is the mean of the dataset, and σ is the standard deviation of all data points.

Second, the normalised time-series dataset, denoted as Z, will be shortened into a time-series of N values, denoted as Z_{short}. This is done by dividing Z into N sub-sequences of equal lengths and extracting the mean value of each-subsequence.

Third, Z_{short} is converted into a string of N letters, with each letter representing a range with equal probability in the normal distribution.

In the context of this research, N is set as 7, i.e., the letters from ‘a’ to ‘g’, and the letter ‘g’ represents the peaks.

References

[1] A. Whiteman, D. Akande, N. Elhassan, G. Escamilla, I. Ahmed, Renewable Energy Statistics 2023, Tech. Rep., International Renewable Energy Agency, 2023, URL https://mc-cd8320d4-36a1-40ac-83cc-3389-cdn-endpoint.azureedge.net/-/media/Files/IRENA/Agency/Publication/2023/Jul/IRENA_Renewable_energy_statistics_2023.pdf?rev=7b2f44c294b84cad9a27fc24949d2134.

[2] Y. Abdelilah, A.A. Báscones, H. Bahar, P. Bojek, F. Briens, J. Criswell, L.L. Laura Mari Martinez, Renewables 2023: Analysis and forecast to 2028, Tech. Rep., International Energy Agency, 2024, URL https://iea.blob.core.windows.net/assets/96d66a8b-d502-476b-ba94-54ffda84cf72/Renewables_2023.pdf.

[3] UK Government, Solar Photovoltaics Deployment, Tech. Rep., Department for Business, Energy & Industrial Strategy, 2024, URL https://www.gov.uk/government/statistics/solar-photovoltaics-deployment.

[4] L. Waswa, M.J. Chihota, B. Bekker, A probabilistic estimation of PV capacity in distribution networks from aggregated net-load data, IEEE Access 9 (2021) 140358–140371.

[5] O. Huxley, J. Taylor, A. Everard, J. Briggs, K. Tilley, J. Harwood, A. Buckley, The uncertainties involved in measuring national solar photovoltaic electricity generation, Renew. Sustain. Energy Rev. 156 (2022) http://dx.doi.org/10.1016/j.rser.2021.112000, URL https://www.sciencedirect.com/science/article/pii/S1364032121012636.

[6] H. Bliss, What is overvoltage?, 2024, https://www.aboutmechanics.com/what-is-overvoltage.htm. (Accessed 1 March 2024).

[7] J. Parker, Reverse power flow, 2024, http://www.hvpower.co.nz/TechnicalLibrary/A-Eberle/Reverse_Power_Flow.pdf. (Accessed 1 March 2024).

[8] X. Zhang, S. Grijalva, A data-driven approach for detection and estimation of residential PV installations, IEEE Trans. Smart Grid 7 (5) (2016) 2477–2485.

[9] K. Li, F. Wang, Z. Mi, M. Fotuhi-Firuzabad, N. Duić, T. Wang, Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation, Appl. Energy 253 (2019) 113595.

[10] B.C. Erdener, C. Feng, K. Doubleday, A. Florita, B.-M. Hodge, A review of behind-the-meter solar forecasting, Renew. Sustain. Energy Rev. 160 (2022) 112224, http://dx.doi.org/10.1016/j.rser.2022.112224, URL https://www.sciencedirect.com/science/article/pii/S1364032122001472.

[11] T. Zhong, Z. Zhang, M. Chen, K. Zhang, Z. Zhou, R. Zhu, Y. Wang, G. Lü, J. Yan, A city-scale estimation of rooftop solar photovoltaic potential based on deep learning, Appl. Energy 298 (2021) 117132, http://dx.doi.org/10.1016/j.apenergy.2021.117132, URL https://www.sciencedirect.com/science/article/pii/S0306261921005729.

[12] S. Krapf, N. Kemmerzell, S. Khawaja Haseeb Uddin, M. Hack Vázquez, F. Netzler, M. Lienkamp, Towards scalable economic photovoltaic potential analysis using aerial images and deep learning, Energies 14 (13) (2021) http://dx.doi.org/10.3390/en14133800, URL https://www.mdpi.com/1996-1073/14/13/3800.

[13] H. Ren, C. Xu, Z. Ma, Y. Sun, A novel 3D-geographic information system and deep learning integrated approach for high-accuracy building rooftop solar energy potential characterization of high-density cities, Appl. Energy 306 (2022) 117985, http://dx.doi.org/10.1016/j.apenergy.2021.117985, URL https://www.sciencedirect.com/science/article/pii/S0306261921012885.

[14] K. Pan, Z. Chen, C.S. Lai, C. Xie, D. Wang, X. Li, Z. Zhao, N. Tong, L.L. Lai, An unsupervised data-driven approach for behind-the-meter photovoltaic power generation disaggregation, Appl. Energy 309 (2022) 118450, http://dx.doi.org/10.1016/j.apenergy.2021.118450, URL https://www.sciencedirect.com/science/article/pii/S0306261921016755.

[15] J. Wang, W. Zheng, Z. Li, Detection and estimation of behind-the-meter photovoltaic generation based on smart meter data analytics, Electr. J. 35 (5) (2022) 107132, http://dx.doi.org/10.1016/j.tej.2022.107132, URL https://www.sciencedirect.com/science/article/pii/S1040619022000586 Behind the meter strategies for enhancing the electricity grid resilience, reliability, economics, sustainability, and security.

[16] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[17] C.C. Liu, H. Chen, J. Shi, L. Chen, Self-supervised learning method for consumer-level behind-the-meter PV estimation, Appl. Energy 326 (2022) 119961, http://dx.doi.org/10.1016/j.apenergy.2022.119961, URL https://www.sciencedirect.com/science/article/pii/S0306261922012181.

[18] X.-Y. Zhang, C. Watkins, S. Kuenzel, Multi-quantile recurrent neural network for feeder-level probabilistic energy disaggregation considering roof-top solar energy, Eng. Appl. Artif. Intell. 110 (2022) 104707.

[19] Y. Kiguchi, Y. Heo, M. Weeks, R. Choudhary, Predicting intra-day load profiles under time-of-use tariffs using smart meter data, Energy 173 (2019) 959–970, http://dx.doi.org/10.1016/j.energy.2019.01.037, URL https://www.sciencedirect.com/science/article/pii/S0360544219300398.

[20] V.Z. Castillo, H.-S. de Boer, R.M. Muñoz, D.E. Gernaat, R. Benders, D. van Vuuren, Future global electricity demand load curves, Energy 258 (2022) 124741, http://dx.doi.org/10.1016/j.energy.2022.124741, URL https://www.sciencedirect.com/science/article/pii/S0360544222016449.

[21] L.P.C. Do, K.-H. Lin, P. Molnár, Electricity consumption modelling: A case of Germany, Econ. Model. 55 (2016) 92–101, http://dx.doi.org/10.1016/j.econmod.2016.02.010, URL https://www.sciencedirect.com/science/article/pii/S0264999316300219.

[22] H. Cho, Y. Goude, X. Brossat, Q. Yao, Modeling and forecasting daily electricity load curves: A hybrid approach, J. Amer. Statist. Assoc. 108 (501) (2013) 7–21, http://dx.doi.org/10.1080/01621459.2012.722900.

[23] Met Office, When does summer start? 2024, https://www.metoffice.gov.uk/weather/learn-about/weather/seasons/summer/when-does-summer-start. (Accessed 27 February 2024).

[24] M. Hu, D. Ge, R. Telford, B. Stephen, D.C. Wallom, Classification and characterization of intra-day load curves of PV and non-PV households using interpretable feature extraction and feature-based clustering, Sustain. Cities Soc. 75 (2021) 103380.

[25] C. Gavin, Seasonal variations in electricity demand, Tech. Rep., Electricity Statistics, 2014, URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/295225/Seasonal_variations_in_electricity_demand.pdf.

[26] Britannica, Numerical measures, 2022, https://www.britannica.com/science/statistics/Numerical-measures. (Accessed 27 October 2022).

[27] J. Lin, E. Keogh, L. Wei, S. Lonardi, Experiencing SAX: a novel symbolic representation of time series, Data Min. Knowl. Disc. 15 (2) (2007) 107–144.

[28] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.

[29] J. Heaton, Introduction to Neural Networks with Java, Heaton Research, Inc., 2008.

[30] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems, " O'Reilly Media, Inc.", 2019.

[31] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.

[32] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[33] scikit-learn developers, 3.2. Tuning the hyper-parameters of an estimator, 2024, https://scikit-learn.org/stable/modules/grid_search.html. (Accessed 1 March 2024).

[34] J. Brownlee, Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions, Machine Learning Mastery, 2018.

[35] European Commission, Photovoltatic geographical information system (PVGIS), 2022, https://re.jrc.ec.europa.eu/pvg_tools/en/#api_5.1. (Accessed 1 March 2024).

[36] S. Adams, personal communication, 2022.

[37] Energy Saving Trust, A comprehensive guide to solar panels, 2022, https://energysavingtrust.org.uk/advice/solar-panels/. (Accessed 27 October 2022).

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[39] T. Gill, How many people have solar panels in the UK?, 2024, https://www.theecoexperts.co.uk/solar-panels/popularity-of-solar-power. (Accessed 1 March 2024).

[40] Qualtrics, Interpreting residual plots to improve your regression, 2024, https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/. (Accessed 1 March 2024).

[41] C. Dong, B. Sigrin, G. Brinkman, Forecasting residential solar photovoltaic deployment in California, Technol. Forecast. Soc. Change 117 (2017) 251–265.

[42] A. De Myttenaere, B. Golden, B. Le Grand, F. Rossi, Mean absolute percentage error for regression models, Neurocomputing 192 (2016) 38–48.

[43] pv magazine international, Residential battery inventories soaring in Europe, says S&P Global, 2023, URL https://www.pv-magazine.com/2023/10/13/residential-battery-inventories-soaring-in-europe-says-sp-global/.

[44] R. Corson, R. Regan, S. Carlson, Implementing energy storage for peak-load shifting, 2014, https://www.csemag.com/articles/implementing-energy-storage-for-peak-load-shifting/. (Accessed 1 March 2024).

[45] S. Alsauskas, E. Connelly, A. Daou, A. Gouy, M. Huismans, H. Kim, J.L. Marois, S. McDonagh, A. Petropoulos, J. Teter, Global EV Outlook 2023, Tech. Rep., International Energy Agency, 2023, URL https://www.iea.org/reports/global-ev-outlook-2023.

[46] H. Engel, R. Hensley, S. Knupfer, S. Sahdev, The Potential Impact of Electric Vehicles on Global Energy Systems, Tech. Rep., McKinsey & Company, 2018, URL https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/the-potential-impact-of-electric-vehicles-on-global-energy-systems.

[47] B. Anderson, P. James, Covid-19 lockdown: impacts on GB electricity demand and CO2 emissions, Build. Cities 2 (1) (2021) 134–149, http://dx.doi.org/10.5334/bc.77, URL https://journal-buildingscities.org/articles/10.5334/bc.77/.

[48] K. Desen, M. Parzen, A. Kiprakis, Impact of the COVID-19 lockdown on the electricity system of great britain: A study on energy demand, generation, pricing and grid stability, Energies 635 (2021) 14, http://dx.doi.org/10.3390/en14030635.

[49] Grid at work, 4 Ways Lockdown Life Affected UK Electricity Use, Tech. Rep., National Grid, 2020, URL https://www.nationalgrid.com/uk/stories/grid-at-work-stories/4-ways-lockdown-life-affected-uk-electricity-use.