CrossMark

Predicting key features of a substation without monitoring

T. E. Lee

Correspondence:
tamsin.e.lee@gmail.com
The Mathematical Institute,
University of Oxford, Woodstock
Road, Oxford, UK

Abstract

Understanding domestic electricity consumption behaviour is very important for planners and operators of electrical networks. Distribution Network Operators (DNOs) are installing monitoring devices at the low-voltage (LV) substation level. However, monitoring devices need to be allocated with consideration. This paper seeks to quantify the relationship between key information from a substation electricity profile, and the type of properties it supplies. Often the number of properties is the most significant predictor for key substation features. Additionally, the proportion of properties which are small-to-medium enterprises (SMEs) affects nearly all key features considered: from the overall load mean and variance, to the annual trend, to peak behaviour. Knowing substation key features is of direct use, for example where to prioritise monitoring. However we also demonstrate how these key features can be used within a so-called 'buddying' framework to estimate the load every half-hour for individual properties.

Keywords: Low-voltage network, Generalised-additive model, Buddying, Genetic algorithm

Introduction

Over the next few decades, electricity demand on the low-voltage network is expected to increase and become more irregular with the uptake of low carbon technologies such as electric vehicles and photovoltaics [1]. The grids' changing demands mean that Distribution Network Operators (DNOs) need to be able to better predict electricity usage and evaluate the stability of their LV networks. Network models enable load flow analysis, and validations of potential energy reduction schemes such as demand side response or energy storage devices [2].

So across the UK DNOs, with support from the regulatory body OFGEM, are researching into strategies which will ensure electricity continues to be supplied safely and reliably, and in a manner that has minimal impact upon the customer [3]. Before monitoring devices, low-voltage electricity customer demands could be estimated through the after diversity maximum demand (ADMD) procedure, or aggregates estimated using the specifications of ACE Report No. 49 (ACE49) [4, 5]. Perhaps the most common estimates for individual load on the low-voltage are provided by Elexon [6], which only distinguishes between households with overnight storage heaters and those without, providing a smoothed approximation. However, for use within a network model for power analysis, the volatility and peak demands are required. So to

match evolving electricity consumption on the low-voltage network, new methods are being developed to predict the electricity load of low-voltage customers, both on an individual level and aggregate level, such as [7–9].

A key development in monitoring is Smart Metering. The UK expects to complete their smart meter roll out by 2020, which potentially provides half-hourly load readings for households. Each household represents a separately metered point on the network, and is allocated a unique meter point administration number (MPAN). Smart meter data helps DNOs assess the LV network headroom [10], the effect of energy storage [8], and the network renewable capacity [11]. However, a DNO is charged for smart meter data [1] so approximations to half-hourly household demand are required, which is a not-trivial task since load on this individual level is volatile.

Alternatively, a DNO can install monitoring devices at LV substations, which measures the aggregate load every half hour for all households on the substation, which may be up to 500 households. However, ideally, this monitoring would be kept to a minimum, and key electricity behaviour features would be available without monitoring. This paper seeks to quantify the relationship between key information from substation monitoring, such as the maximum demand, or the time of the peak load, with information available without monitoring, such as the number of domestic or small-to-medium enterprises (SMEs) a substation supplies.

The findings of this paper two-fold. Firstly, we show that we can estimate key electricity demand features at the aggregate level. This information alone helps DNOs plan and manage, for example, knowing the average and peak load at a substation enables a DNO to make informed decisions regarding load management via storage devices [8]. Secondly, we demonstrate that this method can be used within a so-called 'buddying framework' [12] to provide a realistic half-hourly load profile for a single household.

Method

The data is at the half-hourly resolution for 135 substations, from Monday, 19th May 2014 up to Monday, 16th May 2016, precisely 104 weeks. Each substation provides electricity for 2 to 402 MPANs. Monday, 19th May 2014 (00:00) to Sunday, 17th May 2015 (23:30) is referred to as Year 1, and Monday, 18th May 2015 (00:00) to Sunday, 15th May 2016 (23:30) is referred to as Year 2. Each year has 17,472 half-hours. The substations are located in Bracknell, UK, where data has been collected as part of the Thames Valley Vision (TVV) project. Bracknell is a city in the South-East UK, and is home to many large companies and the local network is typical of much of Britain's network, and therefore the lessons learned here can quickly to applied nationwide [13].

Regarding the predictor variables, we select as simple and straightforward a set of predictor variables as possible, relying only upon information already available to a DNO. For each substation we categorise the connected MPANs based upon their profile class. Standard, domestic households are profile class 1, domestic households with overnight storage heaters are profile class 2, and small-to-medium enterprises (SMEs) are profile class 3 or above. For simplicity, we group all SMEs together. Since photovoltaic generation greatly affect electricity load behaviour, and is also information currently available to a DNO, we further note the number of generation MPANs on each substation, where each generation MPAN relates to a separate, secondary, MPAN allocated for a property with, say, solar panels. Lastly, for a substation without

monitoring, a DNO has access to its' maximum demand indicator (MDI) reading. This value is the maximum demand in an instant, so is greater than the maximum at the half-hour resolution. Since we will be interested in predicting peak behaviour, the MDI is likely to be an important predictor. The MDI reading used here is for the period 1st January 2014 to 31st December 2015. We consider the proportion of MPANs in each of these categories, along with the number of MPANs, as possible predictor variables, see (1).
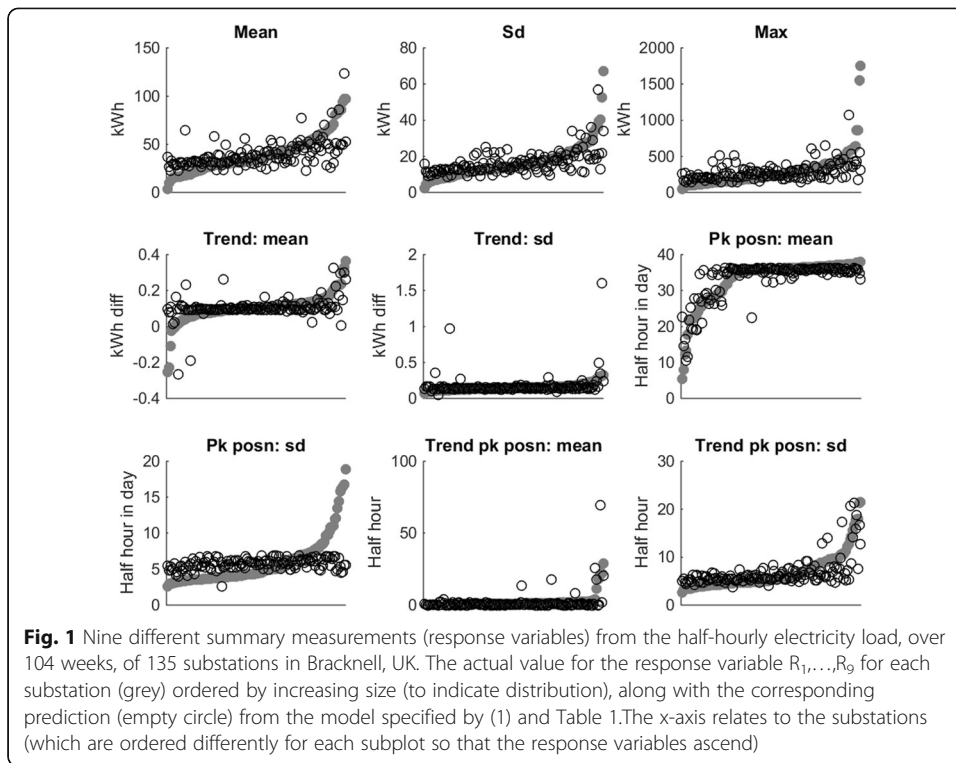
We consider nine features, that is nine response variables R_i, $i = 1,...,9$, which a DNO would like to estimate for substations without monitoring. These response variables can be classed into three primary areas of interest to a DNO: general features, annual trend and peak behaviour. Nonetheless, this method could be applied to predict secondary features such as the ratio of weekday load to weekend load. The first three response variables refer to general features:- R_1: Mean load per half-hour over the 2 year period; R_2: Standard deviation over the 2 year period; R_3: The maximum load (in a given half hour) over the 2 year period. To inform long term decisions, a DNO needs to predict the annual trend, thus the next two response variables are:- R_4: The normalised difference from Year 2 to Year 1; R_5: The standard deviation between Year 1 and Year 2. The network is under the most stress during peak load, so understanding peak behaviour is especially important, so the last four response variables are:- R_6: The mean time of the daily peak over the 2 year period; R_7: The standard deviation in the daily peak position over the 2 year period; R_8: The mean peak value; and R_9: The standard deviation of the peak value.

We use the generalised additive model (GAM) to estimate each response variable R_i, $i = 1,...,9$,

$$R_i = f_1(\text{No.MPAN}) + f_2(\text{Prop.Gen}) + f_3(\text{Prop.PC1}) + f_4(\text{Prop.PC2}) \\ + f_5(\text{Prop.SME}) + f_6(\text{MDI}), \qquad (1)$$

where the functions f_1, to f_6 are smooth functions estimated using a thin plate regression spline. A GAM is flexible to several predictor variables, and enables one to easily interpret which variables are significant. An alternative model would be multiple linear regression, however as this assumes a linear relationship between the response and predictors, it may be unsuitable. Whereas, if the relationship is in fact linear, the GAM model will reduce to the linear model, making a GAM model the most general choice of model. A DNO would prefer a model that requires little information, so if a particular response variable is found to be insignificant for a given response variable, it is preferable to identify this and remove the insignificant predictor from the model. The GAM model is implemented in R using the mgcv R package [14], which also provides a *p*-value for each predictor. For a given R_i, should the *p*-value be greater than 0.05 for a predictor, one can state that the predictor is not statistically significant for the key feature in question, thus the corresponding function f is set to zero. The response variables R_1, R_2, R_3, R_5, R_6, R_7, and R_9 are strictly positive (Fig. 1), so we assume the response follows a Gamma distribution. For R_4, which appears symmetrical, we assume a Gaussian distribution, and R_8, which is skewed, we assume a log Gaussian distribution, see Fig. 1.

To predict the response variables for a single substation s, the model is run without s, then the response variable for s is estimated (leave-one-out analysis). The estimations are compared to the actual values using a root-mean squared error (RMSE). To

Fig. 1 Nine different summary measurements (response variables) from the half-hourly electricity load, over 104 weeks, of 135 substations in Bracknell, UK. The actual value for the response variable $R_1, \ldots, R_9$ for each substation (grey) ordered by increasing size (to indicate distribution), along with the corresponding prediction (empty circle) from the model specified by (1) and Table 1. The x-axis relates to the substations (which are ordered differently for each subplot so that the response variables ascend)

compare the models for the different key features to each other, the RMSE is standardised by dividing it by the mean for each response variable. Without this normalisation, errors for features like the mean difference between years (which is small) will appear smaller than for features like the maximum load (which is large). It also enables a comparison between features that are measured by half hour with features that are measured in kWh.

Results and discussions

The response variables, and the corresponding statistically significant variables, with the *p*-value, are provided in Table 1. The smoothness parameters that control the trade-off between fit and smoothness, is selected by minimisation of the generalised

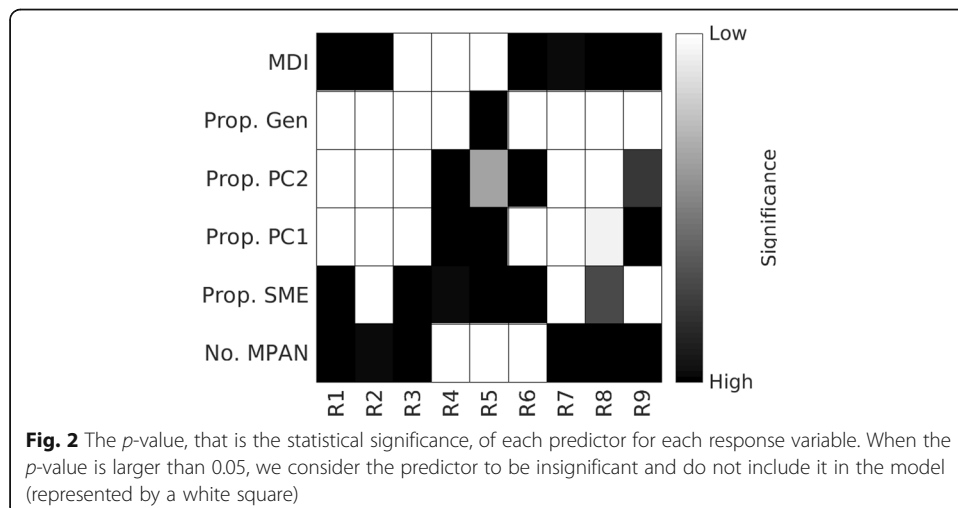**Table 1** The error (RMSE) from the model predictions

R	Key feature (unit)	RMSE	Stnd. RMSE	GCV
6	Peak: mean position (half hour in day)	3.811	0.115	0.0080
8	Peak: mean value (kWh)	21.375	0.310	0.1064
1	Overall mean (kWh)	15.677	0.397	0.1373
2	Overall standard deviation (kWh)	6.751	0.409	0.1434
7	Peak: standard deviation of position (half hour in day)	3.054	0.530	0.1856
4	Trend: mean difference between years (kWh)	0.076	0.733	0.0033
9	Peak: standard deviation of value (kWh)	16.662	0.772	0.1805
3	Overall maximum (kWh)	224.21	0.786	0.2585
5	Trend: std. dev of diff. Between years (kWh)	0.141	0.928	0.0403

The standard RMSE error allows comparison across the nine models. The table is ordered by performance, that is the response variable most accurately predicted is first. A small GCV score means a better fit of the model

cross validation score (GCV) [15]. These scores are also provided in Table 1, where a smaller GCV value indicates a better fit of the model.

The standardised RMSE suggests that peak mean position R_6 and value R_8 are the easiest to predict (Table 1), which is good for a DNO. The RMSE of 3.811 for peak mean position R_6 indicates that on average, the peak is only 2 h incorrect. However, as 90 of the 135 substations (67%) have a peak mean position R_6 between 5.30 pm (half hour 35), and the maximum peak mean position of 7:02 pm (half hour 38.08), this RMSE is perhaps unexpected (Fig. 1). The standard deviation of the year to year difference R_5 is the most difficult to predict, but as there is little variation in R_5, this is of little concern. However, the maximum demand for a given half-hour R_3 varies considerably, and the model struggles to capture the high users (Table 1 and Fig. 1). Surprisingly the MDI reading is not significantly related to the maximum R_3 (Fig. 2). This may be because the MDI reading is for a different time period. The two outliers for R_3 are not unusual substations, they supply 134 and 133 MPANs, nearly all of which are profile class 1 (standard domestic), and none of which are SMEs. The mean loads for these two outliers are 85KWh and 57KWh - nothing extreme that would explain their extreme maximum demand. Moreover, these two outliers are not the same outliers in R_9, the standard deviation of the peak value. These two outliers supply a larger number of MPANs, 203 and 304 which would indicate that there would be less variance around their peak behaviour, not more. The model also struggles to capture the low mean consumption R_1 (Fig. 1), however, as the values are small, and the load is overestimated, this inaccuracy is of little concern to a DNO. The proportion of generation MPANs on a substation only affects the standard deviation of the annual trend R_5 (Fig. 2).

Unsurprisingly, the peak position R_6 depends on the proportion of MPANs which have overnight storage heaters. However, so does the annual trend behaviour - both the mean R_4 and standard deviation R_5. For the majority of substations, the time of the daily peak R_7 does not vary much. Nonetheless, for the few substations which do present a large standard deviation in peak position, the model cannot accurately predict this behaviour (Fig. 1).
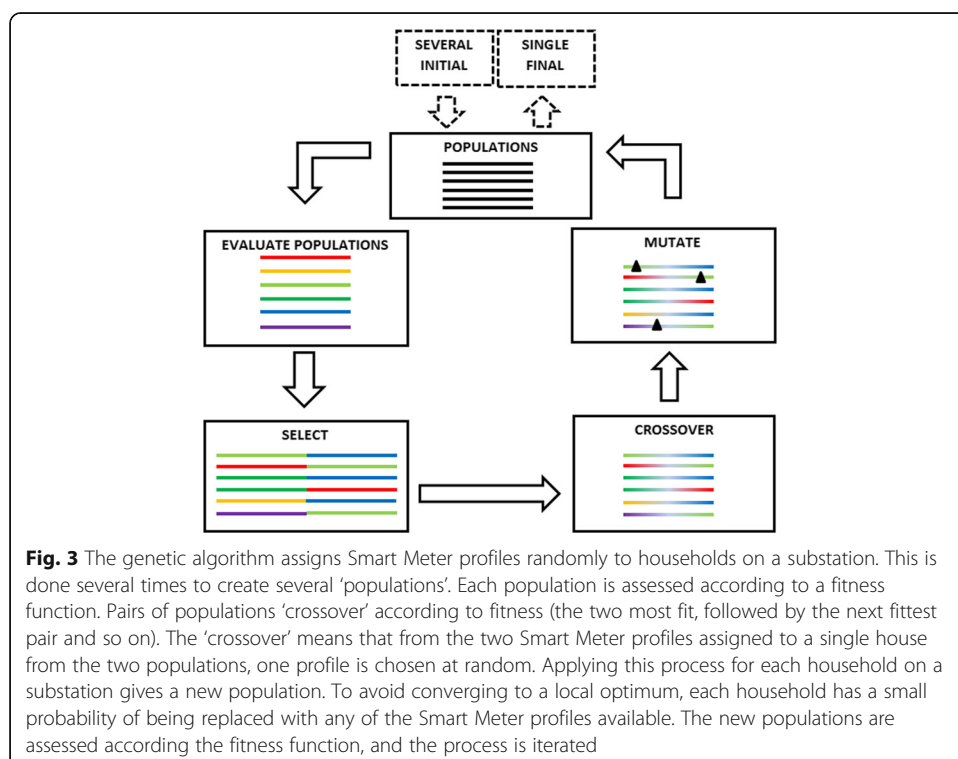

Fig. 2 The *p*-value, that is the statistical significance, of each predictor for each response variable. When the *p*-value is larger than 0.05, we consider the predictor to be insignificant and do not include it in the model (represented by a white square)

Application to buddying

In this paper we refer to a 'smart meter profile' as the reading of the electricity load of a property every 30 min. Buddying is a process which assigns a smart meter profile to a household that does not have a smart meter. Therefore, it is an important tool for a DNO, since even if smart meter data can be purchased, it is likely to be expensive [16]. However, DNOs can install high resolution monitoring at the substation level and thus understand the aggregated demand, and potentially infer the customer demands [12]. Using a genetic algorithm, Giasemidis et al. [12] assign a smart meter profile to domestic customers based on demand at the substation, and households' quarterly meter readings. Their genetic algorithm assigns properties on a substation to a particular smart meter profiles in a manner that minimises the difference between the estimated aggregated demand, and actual aggregated demand (from substation monitoring), and it also minimises the difference between the actual daily load (from quarterly meter readings) and the daily load of the assigned smart meter profile. The authors explore different weightings applied to these two measures. A diagram to illustrate the genetic algorithm is provided in Fig. 3.

To avoid monitoring at the substation, and to avoid relying upon (often inaccurate) quarterly meter readings, we propose a genetic algorithm which uses the predicted key features instead. We use smart meter data for 223 domestic properties over the same time period used for the substation data. The domestic customers are located in Bracknell, UK, where data has been collected as part of the Thames Valley Vision (TVV) project.

The buddying algorithm experiments with different combinations of Smart Meter profiles assigned to one substation. Note that one of the 223 Smart Meter profiles can

Fig. 3 The genetic algorithm assigns Smart Meter profiles randomly to households on a substation. This is done several times to create several 'populations'. Each population is assessed according to a fitness function. Pairs of populations 'crossover' according to fitness (the two most fit, followed by the next fittest pair and so on). The 'crossover' means that from the two Smart Meter profiles assigned to a single house from the two populations, one profile is chosen at random. Applying this process for each household on a substation gives a new population. To avoid converging to a local optimum, each household has a small probability of being replaced with any of the Smart Meter profiles available. The new populations are assessed according the fitness function, and the process is iterated

be assigned to several households on a single substation, that is, we are essentially grouping households into one of 223 categories. The aggregate load from this estimated population is calculated and compared to the actual load at the substation. The estimate and actual load is only compared at two points (whereas [12] aim to match every half hour over several weeks). That is, the fitness function proposed here minimises the mean squared error (RMSE) of the estimated mean peak value, and the mean half-hourly load,

$$F = w|P\text{-}R_8| + (1\text{-}w)|A\text{-}R_1| \tag{2}$$

where P is the mean daily peak load from the aggregated profiles, and A is the half-hourly average load from the aggregated profiles. We experiment with different weighting assigned to peak versus mean, w = 0,0.1,…,1, and find that w = 0.9 (most weighting on matching mean peak load), is optimum. However, to three decimal places there is no difference in the median RMSE for w = 0.6, 0.7, 0.8, 0.9 and 1 (median RMSE of 0.011, see Fig. 4).

As a comparison, we also consider a fitness function that minimises the difference for every half-hour over an average day, a calculation only possible with substation monitoring,

$$G = \sqrt{\left\{\sum (L_h\text{-}l_h)^2\right\}} \tag{3}$$

where the summation is over h = 1,..,48, L_h is the average load at half-hour h from the aggregated profiles, and l_h is the actual average load at half-hour h from the substation monitoring. Using fitness function G (Eqn. (3)) instead of F (Eqn. (2)) actually results in a larger error (median RMSE of 0.199, see Fig. 4). This highlights the problems that can occur when using averages to capture peak behaviour.

To reduce computation time, we impose the limit that it is only possible for 10% of properties on a substation to be allocated the same smart meter profile. The genetic algorithm is implemented using genalg, a package for the statistical software R.
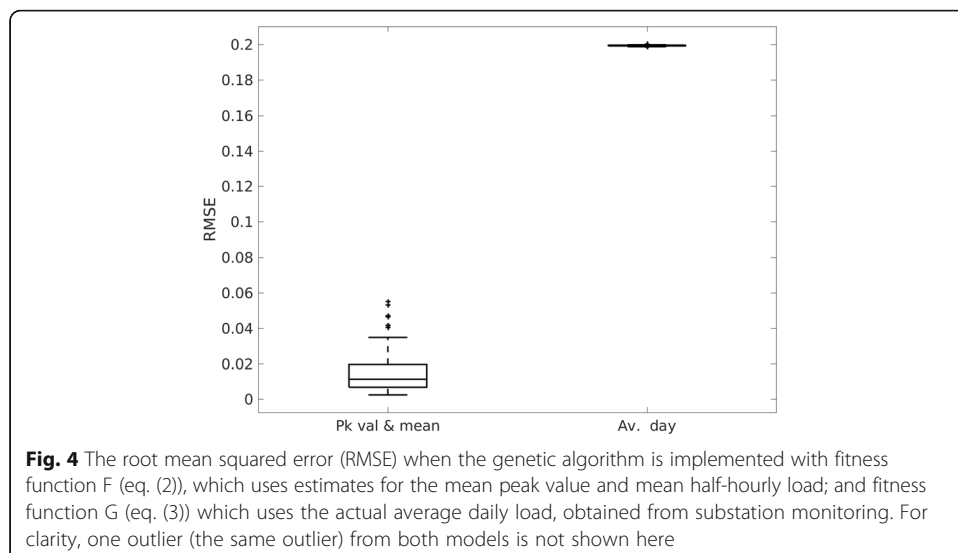

Fig. 4 The root mean squared error (RMSE) when the genetic algorithm is implemented with fitness function F (eq. (2)), which uses estimates for the mean peak value and mean half-hourly load; and fitness function G (eq. (3)) which uses the actual average daily load, obtained from substation monitoring. For clarity, one outlier (the same outlier) from both models is not shown here

Conclusion

The effect on electricity load of the type of MPANs a substation supplies have been quantified, allowing predictions on the overall load, the annual trend, and peak behaviour - with greatest success in predicting mean peak load and position, and overall mean and standard deviation. Generally, the number of MPANs, the MDI reading and the proportion of SMEs are the strongest predictor variables. The importance of the MDI reading is reassuring for DNOs since it confirms the relevance of their current methods. The proportion of profile class 1 customers is required when estimating annual trend features, and the proportion of profile class 2 customers (overnight storage heater customers) is required when estimating the peak position and annual trend. The proportion of generation MPANs has little effect. However, the highest proportion of generation on a substation was only 10%.

For a substation without monitoring, we accurately estimated range of key features of value to a DNO. Therefore the learning gained from current monitoring can be extended to many more substations. This is especially true if the substation does not exhibit behaviour at the extremes, for example, the model did not accurately estimate the maximum load for two substations which exhibit a very high maximum, nor two different substations which have a large range of around the average peak value. However, further analysis showed nothing special about these outliers, suggesting that factors outside of those considered here affect peak demand. Should identifying these outliers be imperative, only monitoring would be able to inform the DNO. Nonetheless, for the majority of features, all substations were predicted with reasonable accuracy.

Lastly, we showed how our predictions for mean peak value, and mean half-hourly load, can be used within a buddying framework [12]. Aggregating across assigned smart meter profiles on a substation, we matched the substation behaviour accurately, even when compared with alternative methods which require substation monitoring. Therefore, the methods presented in this paper would allow a smart meter profile to be assigned to a household without requiring monitoring at the substation, nor requiring quarterly meter readings. A DNO can use these smart meter profiles within a network modelling environment, and make planning and management decisions. For example, electric vehicles and solar panels can be added (see [17, 18]) to the virtual network to inform decisions about reinforcements.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. EA Technology. Assessing the impact of low carbon technologies on great Britain's power distribution networks (2014)
2. Lyons P, Wade N, Jiang T, Taylor P, Hashiesh F, Michel M et al (2015) Design and analysis of electrical energy storage demonstration projects on UK distribution networks. Appl Energy 137:677–691

3. Version seven of the low carbon networks fund governance document, Ofgem, April 2015
4. Richardson I, Thomson M, Infield D, Flifford C (2010) Energy Build 42:1878–1887
5. ENA, Report on statistical method for calculating demands and voltage regulations on LV radial distribution systems (1981) Energy Networks Association
6. ELEXON, Load profiles and their use in electricity settlement, 2013
7. Torriti J, Hassan MG, Leach M (2010) Demand response experience in Europe: policies, programmes and implementation. Energy 35(4):1575–1583
8. Rowe M, Yunusov T, Haben S, Singleton C, Holderbaum W, Potter B (2014) A peak reduction scheduling algorithm for storage devices on the low voltage network. Smart Grid, IEEE Transactions 5(3):2115–2124
9. McDonald J (2008) Adapative intelligent power systems: active distribution networks. Energ Policy 36(12):4346–4351
10. Harrison GP, Wallace AR (2005) Optimal power flow evaluation of distribution network capacity for the connection of distribution generation. Proc Inst Elec Eng Gen Transm Distrib 152:115–122
11. Ochoa L, Dent CJ, Harrison GP (2010) Distribution network capacity assessment: variable DG and active networks. IEEE Trans Power Syst 25:87–95
12. Giasemidis G, Haben S, Lee TE, Singleton C (2017) A genetic algorithm approach for modelling low voltage network demands. Appl Energy 203:463–473
13. http://www.thamesvalleyvision.co.uk/ Accessed 20 Apr 2017
14. Wood S (2006) Generalised additive models: an introduction with R. CRC press, Taylor and Francis Group, USA
15. A. Pierrot and Y. Goude, (2011) Short-term electricity load forecasting with generalised additive models. 16th international conference on intelligent system applications to power systems
16. Technology EA (2014) Assessing the impact of low carbon technologies on Great Britain's power distribution networks https://www.ofgem.gov.uk/publications-and-updates/assessing-impact-low-carbon-technologies-great-britains-power-distribution-networks.
17. Poghosyan A, Greetham DV, Haben S, Lee T (2015) Long term individual load forecast under different electrical vehicles uptake scenarios. Appl Energy 157:699–709
18. Hattam L, Greetham DV (2017) Green neighbourhoods in low voltage networks: measuring impact of electric vehicles and photovoltaics on load profiles. Journal of Modern Power Systems and Clean Energy 5(1):105–116