Forecasting of Residential Electricity Demand using Machine Learning

Feargal Murphy, Loughborough University, Loughborough, Leics. LE11 3TU

A paper for possible publication in *Energies*

Abstract: Accurate electricity demand forecasts are a requirement for decarbonization of the power system. Residential demand profiles are characterised by abrupt temporal changes and are dependent on individual's lifestyle and behaviours. Conventional utility forecasting uses aggregation to average out this individual diversity. This study focusses on short term load forecasting (STLF) for residential consumers. We utilize smart meter (SM) data and apply several forecasting methods at small aggregations ranging from one to several hundred households. We have constructed simple naïve persistence models for benchmarking and compared these with multi-linear regression (MLR) and artificial neural network (ANN) models. We apply these models to generate day-ahead forecasts, using SM datasets from UK (London) and Ireland.

Results suggest that it is inherently difficult to forecast demand at very small aggregations, irrespective of method. Weather variables have less influence than expected. We find that ANN models do not perform better than MLR suggesting that linear terms account for most of the forecastable behaviours. We find that the naïve model often performs as well as more sophisticated models. We conclude that while a useful forecast can be generated for some individual households, aggregations of > 10 households are generally required to achieve an error of 20% or less.

Keywords: Short-Term Load Forecasting, Multi-Linear Regression, Artificial Neural Networks, Naïve Persistence, Smart Meters

1. INTRODUCTION

1.1. Research Context

The expected rapid increase in deployment of variable renewables generation and prosumer 'grid-edge' technologies requires more accurate electricity demand forecasts at higher spatial and temporal granularity compared with traditional utility aggregate forecasts. The main application of a more granular short-term load[i] forecast (STLF) is the requirement to forecast power flow in the grid. This is in the context of diversity of new loads and generators lower down the distribution system (so called 'grid edge') and the requirement to maintain grid stability. More specifically, applications include a) Limiting the effects of network constraints b) Economic scheduling of distributed generating capacity and energy storage including optimization of battery charging and discharge c) Demand response (DR) d) Energy transactions such as peer-to-peer and micro-grid power purchases and dispatch of community energy storage ([1],[2]). Our study focusses on the day ahead forecast horizon, considering that this is critical for optimizing battery storage and DR services.

1.2. Literature Review

Introduction to Short-Term Load Forecasting

A general introduction to classical LF is provided by [3] and [4]. Critical controlling factors include weather, calendar variables, consumer class and electricity pricing.[5] and [6] provide a more recent overview defining STLF as covering the time period from 24 hrs to 2 weeks ahead. Figure 1 illustrates the basic process employed in STLF whereby historic weather and load data are used to train a model. Our study is an ex-post application of trained models, using the measured load values from a test window to assess the accuracy of the forecast model. [7] presents a systematic classification of LF by forecast time horizon, method and dependent variables.

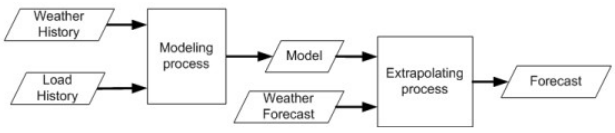

Figure 1 Schematic of Typical STLF Process [5]

Data

[8] and [9], as part of a comprehensive literature review of data analytics in load forecasting, provides a summary of available smart meter (SM) datasets. The LCL (Low Carbon London) [10] and CER (Commission for Energy Regulation, Ireland) datasets were used in this study. [10] provides a detailed description of the LCL trial including approximate

location data (Greater London), consumer recruitment socio-economic classes and key objectives and envisaged applications. [11] describes the recruitment process for the Irish trial. Both trials share some common objectives, namely: a) assess impact of demand stimuli on customers consumption behaviour b) develop experience in implementation of SM grid technology under the general theme of developing low carbon technologies.

STLF Models & Variables

The literature on STLF can broadly be divided into time series models, in which demand is modelled as a function of past observed values, and causal models where demand is modelled as a function of exogenous factors such as weather, calendar or social variables [12]. An alternative classification is provided [3] into a) Time of Day models which include naïve persistence models and b) Dynamic models such as MLR and Autoregressive moving average models in which load is also a factor of recent history and exogenous factors.

Variables to be considered in STLF are enumerated by [4]. These include time and calendar (hour of day, day of week, seasonal), weather (temperature, composite humidity-windchill-temperature) and customer class (residential, commercial). Weekdays adjacent to weekends and holiday may have different characteristics to normal weekdays [13].

The assumption when applying linear regression (LR) is that output is linearly related to input variables and any deviation is caused by observation error. The mathematical foundation for LF regression models is provided by [14] while [5] provides a detailed description of the multi-linear regression (MLR) method, classifying predictors in two categories a) qualitative indicator (dummy) variables such as day of week and hour of day and b) quantitative variables such as temperature. These variables may have interaction effects [13] e.g. load may be higher in evening than morning at similar temperatures. Dummy variables may be used to account for interaction effects that may not be obvious from a direct single variable scatter plot correlation analysis [15].

[12] provide a review of literature on application of ANN in STLF, including an introduction to the multi-layer perceptron (MLP) and common pitfalls such as inappropriate architecture and over fitting. The most common ANN method for STLF has been feed-forward NN with back-propagation [4]. Recent studies have used recurrent and deep learning [16] but care must be taken to avid over-fitting. [17] carried out a thorough examination of hyperparameters used for STLF, finding evidence to support using up to 5 hidden layers in the NN. [18] concluded that there is not a universally 'best' method and that data and business need should determine selection.

Customers from different classes have markedly different typical demand profiles but within the residential class, there is a typical daily shape with morning and evening peaks and overnight lows (Figure 2).

Examples of STLF for Dis-aggregated Residential Loads

A recent set of studies perform STLF for household aggregation levels. [19] have run six traditional forecast models on two clusters from the CER dataset comprising 90 and 230 households, respectively, both on standard tariff. The pool sizes are representative of a typical community micro-grid or virtual power plant (VPP). They find that while regression provides the best average results, neural network (NN) models better predict demand peaks and troughs.[15] looks at highly dis-aggregated data from experiments run over several weeks to months in Germany and USA on 1 and 6 households, respectively. A key conclusion is that sophisticated methods rarely perform significantly better than naive persistence models at individual household level if they are not adapted to embed individual household attributes.

[20] use proprietary data from 180k residential consumers in California, apply ARIMA, ANN and support vector regression (SVR) methods to predict loads from 1-24 hrs ahead. They develop an empirical scaling law ('Aggregation Error Curve'), showing how errors for residential consumers scale with aggregation size. They conclude that error is reduced up to a critical aggregate consumption beyond which there is little improvement.

[16] and [21] provide what seems to be the most accurate forecasting result for STLF of individual households over 1-24hr time scales. [16] have used the CER data to train a deep learning (DL) model and forecast day ahead load values. They benchmark against the performance of more traditional models and illustrate that the DL model gives a substantial improvement in performance of up to 20% when averaged across households. However, while the DL model is trained on the full dataset, the benchmark models are trained on individual households. [21] use probabilistic forecasting and develop an ensemble model which they apply to both aggregate and individual household examples from the CER data. In several of these studies, it is therefore not clear how representative the illustrated household examples are of the more general applicability and accuracy of the respective methods across the full dataset. [22] compare several ANN and DL methods and are one of the few studies which use the LCL data for STLF. They train on subsets of residential customers and find that DL models forecast peak demand more accurately.

A common theme emerging from both the literature and our findings is that 'unpredictable' consumers are difficult to forecast, irrespective of method.

1.3. Aims & Objectives

This study aims to establish what analytic or machine learning forecasting method or combination thereof, generates most accurate day-ahead forecast of residential consumer electricity demand, and what level of household aggregation is required to generate repeatable robust results. The following are key objectives:

- Review existing public domain residential consumer smart meter (SM) datasets and select a subset that satisfy these criteria: a) sufficient households (several hundred+) and load profile diversity b) sufficient measurement duration (1yr+) c) 1 hour sampling interval or less.
- Adapt and develop sample python code to a) analyse electricity consumption patterns b) build forecasting models utilizing naïve persistence, multi-linear regression (MLR) and artificial neural networks (ANN).
- Review and select appropriate error metrics for load forecast evaluation and comparison.
- Identify how forecasting error behaves with household aggregation level.
- Identify if any particular forecast method consistently outperforms others at different aggregation levels
- How does predictability in general changes across locations?

2. METHODOLOGY

2.1. Data

Following our literature & web search, we catalogued public domain electricity demand datasets that meet required criteria for this study (1.3). We selected the LCL (London) and CER (Ireland) datasets since these were the only datasets that satisfy all the listed criteria at the time of search. The LCL data was selected as the primary dataset for our study, while the CER dataset was selected for comparative analysis. While the CER dataset has been used in many STLF studies, the LCL dataset has been under-utilized in STLF studies, at least until very recently [22]. The LCL data comprised a raw 11GB csv file containing data for > 5000 residential customers spanning all 2013. The CER dataset is supplied as a series of zipped text files containing data for > 5000 residential customers who participated in the Consumer Behaviour Trial (CBT) throughout the official test period (1/Jan-31/Dec,2011). Both datasets contain unique customer (meter) identifier, date-time, demand per 30min window in kWh and a tariff identifier. The LCL data has a socioeconomic 'Acorn' [23] category. LCL data is from customers throughout Greater London while there is no location data supplied for the CER data (CBT customers were selected from 'throughout the country' [11]).

Several data cleaning steps were required for both datasets. Data was imported and processed in chunks as the raw files are too large for memory. For our purposes, it is important to exclude users who were part of any demand stimuli trials as these may bias forecasting results. We selected only users in a standard tariff control group from both datasets.

Processing steps included a) removal of customers not on standard tariff b) trimming of data from periods before/after the calendar year of trial c) removal of duplicate rows d) removal of users with data gaps or periods of zero demand e) conversion to kWh consumption for hourly intervals. For the CER data, since most customers had at least a few readings of 0 values, users were only removed only if there were more than 10 such in the calendar year, as otherwise, most users would have been removed from final database. Retained user data was iteratively appended to a Pandas data frame which was then exported to a serialized binary file for subsequent use. Hourly and daily weather data was extracted from the Dark Sky API for a single representative location. Short gaps (< 3 hrs) were interpolated, and data was plotted as a series of annual heatmaps and monthly profiles for quality checking. Data for temperature, wind speed, humidity, cloud cover and apparent temperature were extracted. Daylight length was calculated from forecast sunrise and sunset times for London and Dublin. Since no location data is available for the Irish dataset, we have assumed a Dublin location for weather data extraction.

2.2. Toolkit

All data preparation, analysis and modelling were performed using the Anaconda Python distribution. Code development was done in Jupyter Lab which facilitates multi-notebook editing. A dedicated virtual environment was set up within Anaconda. Functions were written to achieve generalization and efficiency. These have been combined into a project module, which was maintained using PyCharm. This module can be imported into any notebook, thereby facilitating

access by multiple notebooks and future users. Most code was run on a HP H250-G6 laptop with Intel i7 CPU 2.7 GHz processor and 16MB RAM. However, resource intensive modelling was run on the Google Colab cloud platform using the default CPU option. This facilitated a x10 reduction in computation time and allowed up to 12 hr duration virtual machine sessions.

2.3. Data Analysis

Initial data analysis includes plotting of individual and aggregated load profiles (Figure 2). Aggregation of load series was computed by a) user b) hour of year and plotted to assess user and daily/seasonal demand patterns. Aggregation by user is most easily viewed as a histogram (Figure3a). This identifies users with unusually high or low aggregate demand.

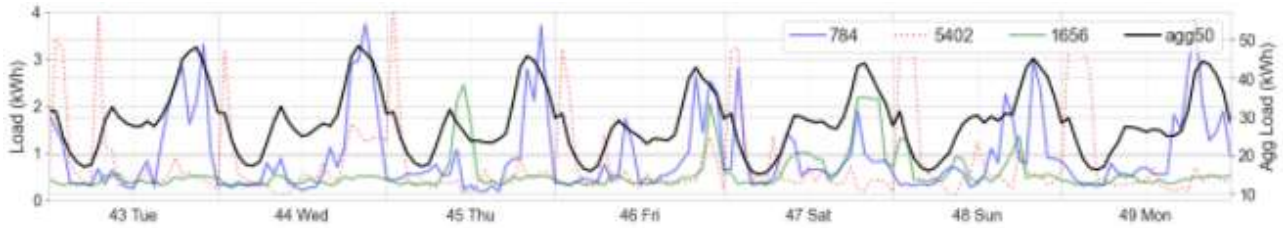

Figure 2 Example Individual and Aggregated (50 user) Load Profiles, Mid-Feb,2013, (LCL Data).

Autocorrelation of load profiles were computed and show the most prevalent periodicities (Figure3b). We expect that demand at time t, l_t, would be similar to demand at the same time on the previous day, $l_{t-24,}$ particularly for weekdays. The autocorrelation identifies a clear 24hr correlation with the maximum at 24hrs lag for most users. The correlation for the aggregate and some users shows a slight increase at a lag of 168 hrs (1 week).

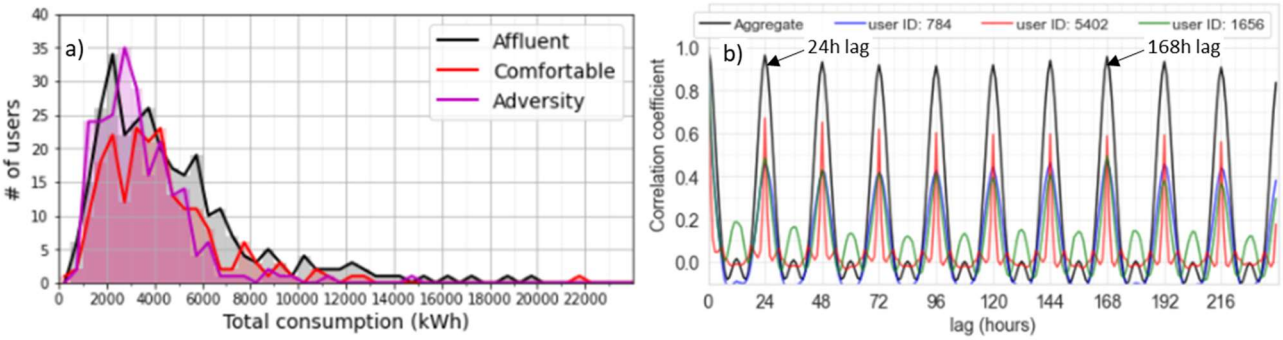

Figure3 a) Histograms of User-Aggregated Load Over 2013 Calendar Year for Major Socio-Economic Categories. b) Autocorrelation of Aggregate and Individual Load Profiles. (LCL data)

Scatter plots of hourly aggregate load against weather variables show what if any, correlation exists over daily and seasonal time scales, and helps to establish which variables may be useful for training regression models (Figure 5). Users with anomalous demand may skew forecast models. To investigate any relation between annual household demand and socioeconomic factors, we formed histograms of annual user aggregate for each Acorn class (Figure3a). There is only minor correlation between median consumption and Acorn group. Load profiles were plotted for high consumption users and show a) normal daily patterns b) such users are split across all 3 groups. Therefore, we only removed 11 users with very low (< 1000 kWh) annual consumption from the LCL database retaining 711 households. The same approach was applied to the CER data, giving a final database with 675 households.

2.4. Models

MLR is a simple, computationally efficient method which captures the influence of quantitative and qualitative factors [5]. As an alternative but computationally more expensive method, we used ANN. We did not use any of the classical time series models such as auto-regressive methods as these are more suited to very short horizons of a few hours [5] whereas our target forecast horizon is 24hrs ahead. Our MLR and ANN models were trained on 6 periods of 5 weeks each, while model performance was tested on the adjacent 2-week windows (Figure 4). Because the persistence model looks back up to 168hrs, a one-week gap is required at start of each period so that test window data is not used for training. This configuration ensures training data spans the full annual seasonal variations in demand (48 weeks). We used weekday and weekend classifiers but did not separate public holidays.

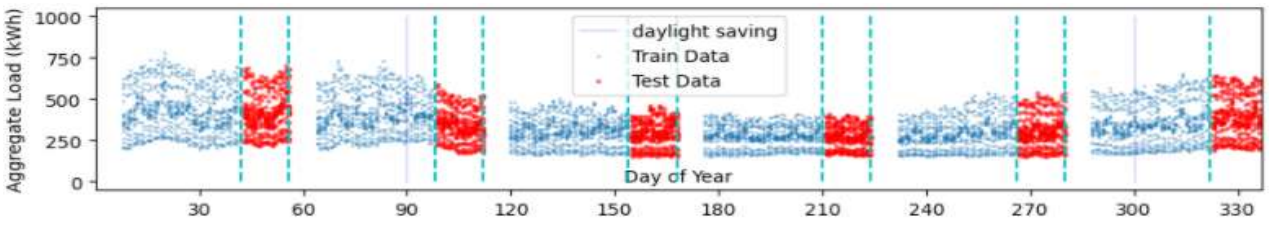

Figure 4 Train and Test Windows with LCL Hourly Aggregate Load.

2.5. Development of Naïve Persistence Model

The 24hr and 168 hr persistence model forecasts are expressed by equations (1) and (2) respectively. In our day classifier, Tues to Fri are assigned to the weekdays class and Sat to Mon are assigned to weekend class. While we expect Tue to Fri to be modelled by equation (1), we intuitively expect Sat, Sun and Mon to be better modelled by equation (2) [13].

$$\hat{l}_t = l_{t-24} \quad (1) \qquad \hat{l}_t = l_{t-168} \quad (2)$$

We formed scatter plots of l_t against l_{t-X} where $X \in \{24, 168\}$, for a) aggregate b) individual users (Figure 7).We computed the linear regression fit and the regression R^2 coefficient to give a quantitative indictor of the persistence model fit. The aggregate case shows a strong linear relationship with $R^2 \sim 0.9$ while individual users generally show poor correlation.

2.6. Development of MLR Model

We plotted aggregate load against several weather/environment variables including temperature, humidity, cloud cover, wind speed, daylight length and sunset time to establish if any correlation. We find that for both the LCL and CER datasets, only temperature, daylight length and sunset times show significant correlation with load (Figure 5).These scatter plots show distinct distributions of load v temperature for different times of day, and load v daylight length for different periods of the year. This is not surprising since a) people respond to temperature differently at different times of day depending on their activities [13] b) we intuitively expect that Autumn and Spring days, having similar daylight length but different average temperatures, will have higher consumption on colder March days compared with warmer September days with similar day length. Following this analysis, we have incorporated 4 daily time bands and 2 annual periods (Jan-June, July-Dec) in the MLR models. We found that a 2nd order polynomial gives the best fit in each band. Although daylight length and sunset show very similar responses, we chose daylight length as the only length of day parameter since sunset times have one hour shifts at start/end of daylight-saving time.

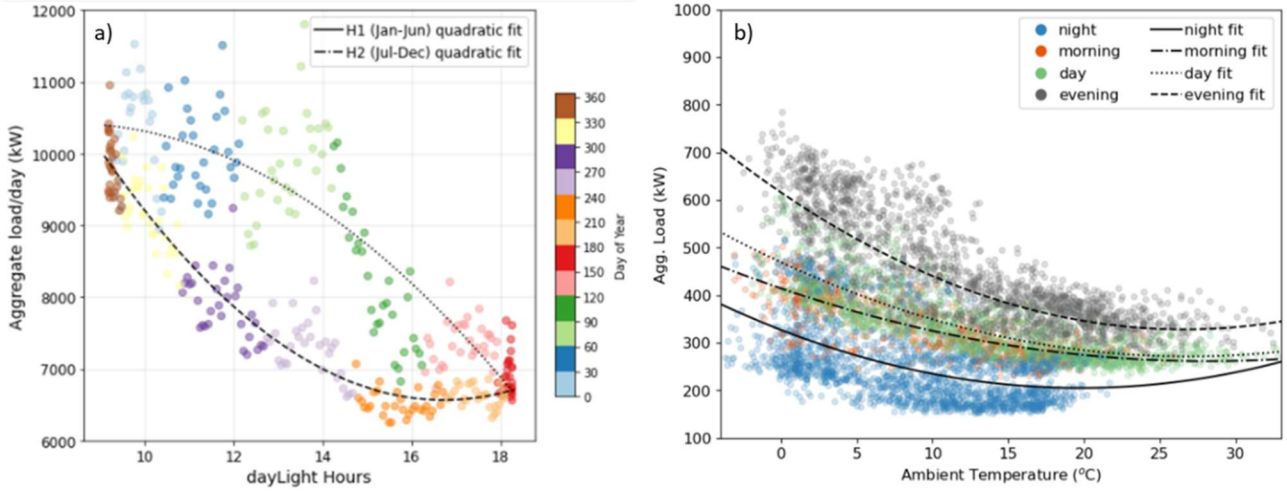

Figure 5 a) Load v Daylight Hours with Quadratic Fit to i) Jan-Jun ii) Jul-Dec Periods. b) Aggregate Load v Temperature with Quadratic Fit for Different Times of Day.

We apply a similar regression model to [5] and [13]. The load, l_t, can be expressed as a linear model of one or more qualitative and quantitative predictor variables:

$$\widehat{l_t} = \sum_{i=1}^{2} DoW_{t,i}\left(\beta_{i,0} + \beta_{i,1} l_{t-24} + \beta_{(2,i)}\, l_{t-168}\right) + \sum_{i=1}^{4} HoD_{t,i}\left(\beta_{3,i} + \beta_{4,i} T_t + \beta_{5,i} T_t^2\right) + \sum_{i=1}^{2} PoY_{t,i}\left(\beta_{6,i} + \beta_{7,i} L_t + \beta_{8,i} L_t^2\right) \quad (3)$$

In equation *(3)*, β terms are the regression coefficients; T and L are quantitative predictor variables temperature and length of day (daylight hours); $DoW_{t,I}$, $HoD_{t,I}$, $PoY_{t,I}$ are qualitative predictor variables corresponding to day of week, time of day and period of year. These are defined in equations (4), *(6)* and *(5)* respectively.

$$[DoW_{t,1}, DoW_{t,2}] = \begin{cases} [1,0] & IF\ day\ is\ Tue, Wed, Thurs, Fri \\ [0,1] & otherwise \end{cases} \quad (4)$$

$$[PoY_{t,1}, PoY_{t,2}] = \begin{cases} [1,0] & IF\ \ day \in Jan\ to\ June \\ [0,1] & IF\ day \in\ July\ to\ Dec \end{cases} \quad (5)$$

$$[HoD_{t,1}, HoD_{t,2}, HoD_{t,3}, HoD_{t,4}] = \begin{cases} [1,0,0,0]\ IF\ t \in\ (23,0,1,2,3,4,5,6) \\ [0,1,0,0]\ IF\ t \in\ (7,8,9) \\ [0,0,1,0]\ IF\ t \in\ (10,11,12,13,14,15) \\ [0,0,1,0\ ]\ IF\ t \in\ (16,17,18,19,20,21,22) \end{cases} \quad (6)$$

Three variants of the MLR model have been explored incorporating 1) quadratic dependence on temperature and daylight length 2) quadratic dependence on temperature and 3) an 'unaware' model with no explicit dependency on either temperature or daylight length. All MLR models incorporate the persistence models and day classifier terms.

2.7. Development of ANN model

Artificial Neural Networks do not require an explicit function defining the relationship between input series and target load series [24]. The basic architecture is shown in Figure 6 and described by equations (7) and (8), consisting of an input layer ($X_1, \ldots, X_T$) and one or more 'hidden layers' (HL) of neurons ($Y_1, \ldots, Y_j$). This is a feed-forward neural net (FFNN) such that outputs from one layer are inputs to next layer and there are no connections between neurons in any given layer. The weights and biases are adjusted during training using gradient decent back-propagation algorithm [24]. The hidden layers allow for non-linearity between input and output. Without these, the ANN would simply be a linear regression. Hyperparameters of the ANN include the number of hidden layers, number of neurons in each hidden layer, the activation function and the learning rate. Normalizing the input data usually prevents computational problems and improves the functionality of training algorithms [22]. We use the minmax scaling method which linearly transforms the value of each variable to between zero and one.

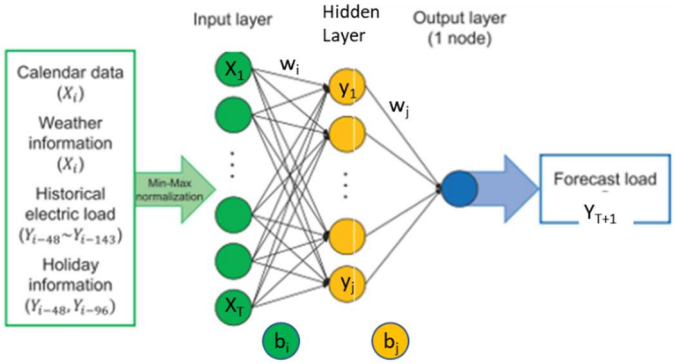

Figure 6 Schematic of STLF Feed Forward NN (Adapted from [18]).

$$y_j = f_1\{\sum_{i=1}^{T} x_i w_i\} + b_i \quad (7) \qquad y_{T+1} = f_2\{\sum_{j=1}^{n} y_j w_j\} + b_j \quad (8)$$

There are 3 stages to generating the ANN forecast: training, validation, and testing. Validation involved holding back some of the training data to optimise parameters. We allocated an 80:20 training / validation data split from the training window data. Once the ANN is trained, it is tested on a separate independent dataset (the 'test' data) to establish whether it is sufficiently generalised to be applied to unseen data. We tested various hyper-parameters using a grid search methodology at the aggregate forecast level (LCL data) since this was relatively quick to compute. Our selected parameters were guided by values used in similar published studies and the values that worked best in our grid-search. We tested variations in number of hidden layers, number of neurons in each hidden layer, the number of epochs and the batch size. For other parameters, we used settings recommended in the literature. A summary of parameters is provided in Table 1. As input series, we used persistence models l_{t-24} and l_{t-168} along with weekday and weekend binary classifiers, temperature and daylight hours. Intuitively, we expect that the non-linear nature of the ANN might allow it to learn patterns that are not captured in the linear MLR model. For example, could it identify patterns where daily load peaks are shifted slightly in

Table 1 ANN Hyperparameters and Values Used in Analysis.

Model Creation		Model Compilation		Model Fitting	
i/p features	dow1,dow2, l(t-24),l(t-168), temp, dayLight	*loss function*	Mean Sq Error	*#epochs*	100
# hidden layers (HL)	1 (LCL), 2 (CER)	*optimizer*	Adam, default learning rate, 0.001	*#batches*	32
# nodes / HL	8 to 30	*metrics*	Mean Sq Error	*validation split*	0.2
Model	Dense (HLs)				
Activation Function	RELU for HL, Linear (none) for o/p layer				
Initialization (weights & biases)	default (glorot_uniform)				

time from day to day i.e. as we might expect for typical residential customers. On this basis, we added loads at time lags > 24hrs e.g. l_{t-25}, l_{t-26} to the input series. We ran comparable configurations for the MLR model including these additional time lags, expecting that we may not see any improvement since the imposed structure is highly linear (Table 2).

2.8. Error Analysis

[24], [25] and [16] introduce the error metrics used in LF. The MAPE (mean average percentage error) metric is the most commonly used as it is a simple relative value, allowing comparisons between datasets, and is accurate provided very small load values are excluded. We computed several metrics including MAPE (9), coefficient of variation (CV), mean average error and root mean square error. We only considered point-based error metrics and decided to use MAPE as the principal metric since it is scale-independent (section 1.2). Although MAPE produces infinite or undefined values if the actual demand is zero or close-to-zero, this disadvantage is minimised by a) using aggregation of series b) defining a load value minimum for error calculations. There are better measures which avoid the double-penalty effect where forecast peaks and troughs are slightly shifted from actual values [26]. However, these are more complex to compute and resource intensive for the large number of forecasts required.

$$MAPE = \frac{100}{n}\sum_{t=1}^{t=t_n}\frac{\hat{l}_t - l_t}{l_t} \qquad (9)$$

2.9. Aggregation Models

We applied our models at a range of aggregation levels from 1 household to all households. We iteratively selected random users in out of sample selection mode, whereby once allocated to a group, a user is excluded from the remaining selection pool. For each N ϵ {1,3,5,10,20,30,40,50,75,100, ALL users}, we formed a set of unique groups of N households until < N households remained in the pool at which point the process jumps to the next N level and the remaining users are discarded. We compute each model forecast and corresponding MAPE error for each aggregation group.

Table 2 Configurations of input series for MLR (left) and ANN (right). For the MLR, only the contributing series are listed as the actual inputs include linear and quadratic terms of these (Section 2).Configurations with same input series are highlighted in same colour.

MLR Input Config	Day of Week Classifier		Load Series						Env Series		Total # Terms in MLR input matrix
	dow1 (Tu-Fr)	dow2 (Sa-Mo)	l(t-24)	l(t-168)	l(t-25)	l(t-26)			Temp. °C	Daylight Hrs	
MLR Config1	✓	✓	✓	✓	x	x			✓	x	6+12
MLR Config2	✓	✓	✓	✓	✓	✓			✓	x	10+12
MLR Config3	✓	✓	✓	✓	x	x			✓	✓	6+18
MLR Config4	✓	✓	✓	✓	✓	✓			✓	✓	10+18

ANN Input Config	Day of Week Classifier		Load Series						Env Series		Total # ANN input series
	dow1 (Tu-Fr)	dow2 (Sa-Mo)	l(t-24)	l(t-168)	l(t-25)	l(t-26)	l(t-27)	l(t-28)	Temp. °C	Daylight Hrs	
ANN Config1	✓	✓	✓	✓	✓	✓	x	x	✓	✓	8
ANN Config2	✓	✓	✓	✓	x	x	x	x	✓	✓	6
ANN Config3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
ANN Config4	*similar to ANN Config2 but a) only Temp, not daylight b) organised as per MLR config1 input matrix with linear and quadratic terms => 6+12 terms*										18

3. RESULTS AND DISCUSSION

3.1. Persistence Model Results

A regression of persistence model forecast is shown in Figure 7 while some individual forecasts are shown in Figure 9. The MAPE errors for the aggregate forecasts (LCL and CER) are given Table 3. A surprising result is that the 24hr persistence model is less accurate than the 1-week model for the Irish dataset. These errors are averaged across weekdays and weekends. Generally, the persistence model is only ~ 1% less accurate than the other models.

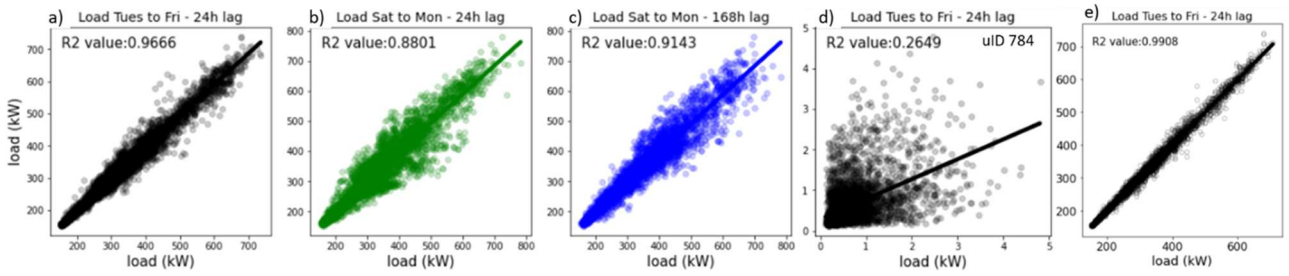

Figure 7 Persistence models for a) 24hr weekdays b) 24hr weekends c) 168hr weekends (all aggregate) and. d) 24hr weekday model for sample individual user e) Similar to a) but where search window is t-X±4 hrs (section 3.1) .

Effect of Window Width on Forecast Errors: Ex-Post Test

We intuitively expect that while typical user demand peaks may not happen at identical times each day, they are likely to fall within a narrow window from day to day. We constructed an ex-post test to examine how the persistence model R^2 error could be reduced by widening the search window i.e. sacrificing temporal accuracy. Instead of regressing the values given by equations (1) and (2) against l_t, a set of load differences ($l_t - l_{t-X-n}$) are constructed for each t, where $X \in \{24,168\}$ and $n \in \{-12,-6,..,0,\ldots,6,12\}$. Iterating over t, for each n value, the l_{t-X-n} value which gives the minimum difference is then assigned as the new l_{t-X} value and regressed against l_t. An R^2 value is computed for each {X,n} and plotted as a histogram (Figure 8a). This test was also run on a suite of household aggregation levels (Figure 8b). These results confirm that for weekends, the t-168 model is better than the t-24 model (Figure 8a) and suggest the potential to predict demand in a ±1 hrs window is considerable better than for a single 1 hr window.

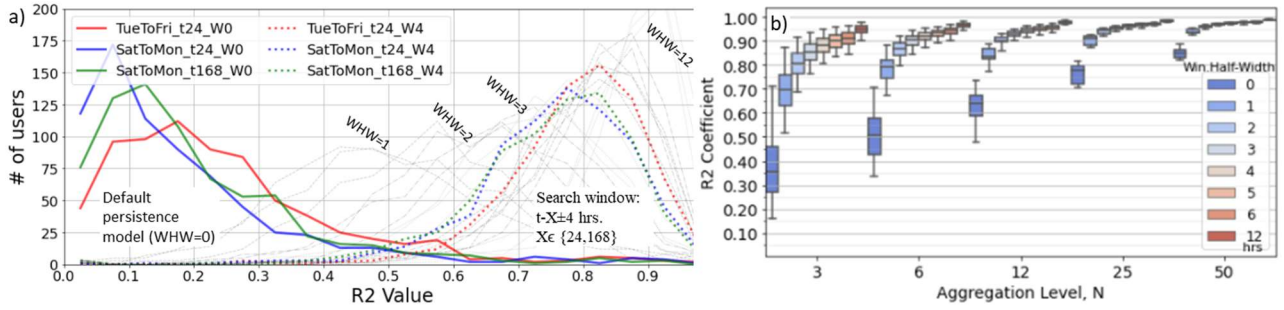

Figure 8 a) Histogram of R^2 error distribution for individual users. Highlighted are the histogram for default persistence models (W0) and the case (W4) where search window half-width is t-X±4 hrs b) Boxplot showing R^2 error distribution for the weekday class at a range of household aggregation levels, N. From left to right, at each N, the search window width varies in range t-X±12 hrs.

3.2. MLR Results

The configurations run for MLR and ANN across all aggregation levels are listed in The MAPE errors for MLR model forecast for the full aggregate of all users are listed in Table 3rs for training and test windows are comparable The addition of daylight hrs does not reduce the error. The *quadratic* models are only marginally better than the *unaware* models for LCL and are slightly worse for CER. This is because daylight length and temperature are already implicitly contained in the persistence model, as both are highly correlated at the forecast time scale. Examples of forecast for each model type is shown in Figure 9. The error distribution by aggregation level for all models is shown as a boxplot Figure 11. The results show that while the forecast works well on an aggregate level, with errors of 5-6%, the median errors at individual and small aggregation levels are much larger e.g. 20% at N=10 and 56% for N=1 for the LCL data.

Table 3 Forecast errors for aggregate models. ANN model results are the mean of 10 iterations (σ = 0.2% for LCL).

Model Name	Input Load Series DC = day classifier	Weather Variables	LCL (London) Test Error	CER (Ireland) Test Error
Persistence t-24			5.87%	7.37%
Persistence t-168			7.03%	6.41%
Unaware	l_{t-24}, l_{t-168}, DC		4.77%	5.76%
QuadTemp, Conf.1	l_{t-24}, l_{t-168} DC	Temperature	4.64%	5.79%
QuadTemp, Conf.2	$l_{t-24}, l_{t-25}, l_{t-26}, l_{t-168}$,DC	Temperature	4.64%	5.71%
QuadTempDLight	$l_{t-24}, l_{t-25}, l_{t-26}, l_{t-168}$, DC	Temp., DayLight	4.68%	5.85%
ANN config1	$l_{t-24}, l_{t-25}, l_{t-26}, l_{t-168}$,DC	Temp., DayLight	5.05%	6.39%
ANN config2	l_{t-24}, l_{t-168} ,DC	Temp., DayLight	4.86%	6.40%
ANN config3	$l_{t-24}, l_{t-25}, l_{t-26}, l_{t-27}, l_{t-28}, l_{t-168}$, DC	Temp., DayLight	4.88%	6.11%

We formed boxplots from the MAPE error distribution for each model (Figure 11). These show that for the LCL data, at aggregation level, N=1, both quadratic and ANN models have similar errors while for N> 1, the error increases from smallest for the MLR quadratic model to maximum error for the 168-hr persistence model. For the CER data, the pattern is similar but a) the ANN model is significantly worse than the quadratic and persistence models at N=1 b) The 168-hr persistence model is better than the 24-hr model for larger N values. This reflects the average result from the aggregate forecast for CER (Table 3).

3.3. ANN results

The ANN training across all aggregation levels took several hrs cloud computing time for each configuration. Hence, only a limited set of configurations could be tested. By comparing forecast profiles across train/test windows and forecast errors for testing and training periods we can establish whether the models are sufficiently generalised. Figure 10 shows that the error is slightly larger for the test windows, particularly for lower aggregation levels which suggests that some

over-training may still be occurring. The theoretical maximum number of trainable parameters in the FFNN is determined by number of inputs, number of HLs and nodes in each HL. It is important that the number of input training samples are at least an order of magnitude greater. Where there is insufficient diversity in the input data, the forecast model may not be sufficiently generalised.

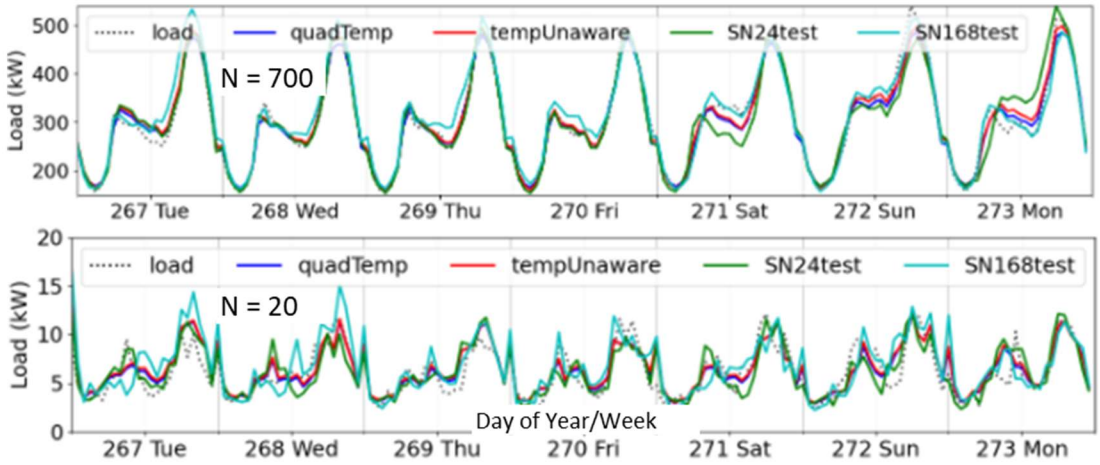

Figure 9 LCL Forecast Examples for MLR and Persistence Models at Aggregation, N= 100, N =20.

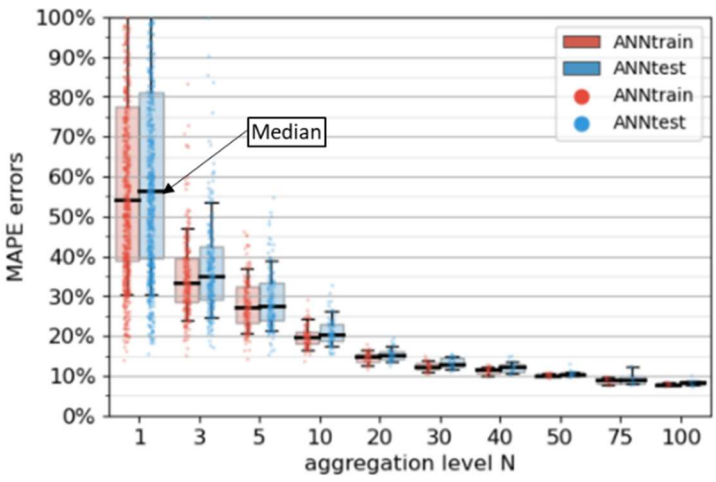

Figure 10 MAPE Errors for TRAIN and TEST Forecasts, LCL. Boxes Represent 2nd and 3d Quartiles, Whiskers show 5 and 95% Ranges.

To establish whether, at low aggregations, (N <= 10), the model performance correlated with households, i.e. do predictable households have low errors for all models and conversely for 'unpredictable' households, we regressed model errors. The persistence model errors are highly corelated and the quadratic and unaware model errors also have a high correlation with the persistence model errors. Figure 12a illustrates how MAPE errors correlate between ANN and quadratic models for each tercile of 24-hr persistence model errors (only aggregations 1 to 10 are shown). We plotted two end member groups (N=3) a) A group with high ANN errors and low errors for other models b) A group with low ANN error but larger error for other models (Figure 12b, 12c). Although the profiles are only shown for a week in winter, they illustrate that group 167 which has higher ANN model errors, has low consumption whereas group 49 which has lower ANN model errors has much higher consumption. This illustrates a challenge with using MAPE as a metric where the denominator in equation (9) is small.

We extracted 25%, 50% and 75% percentile bands at each aggregation level (Figure 13). This facilitates comparison between results for different configurations and locations. Surprisingly, we find that there is very little difference between the MLR and ANN errors at each aggregation level. If anything, ANN performs slightly worse than MLR, particularly for the CER data. We find that there is essentially no significant difference between the various configurations listed in Table 3 . There is a slight improvement in the ANN forecast error when we add additional load series (e.g. l_{t-25} , l_{t-26}) but the MLR result is still marginally better at small aggregations. Although this pattern is repeated for LCL and CER results, the differences are unlikely to be statistically significant. These observations do suggest that given the same inputs, the ANN and MLR give broadly similar forecast accuracy and the non-linear capacity of the ANN is not generating more accurate forecasts in this study. When we supply the ANN with additional inputs, there is a possibility that our data size is not large or diverse enough to see a benefit? The main differences arise a) from different aggregation levels b) location (the Irish dataset shows much larger errors). Our results confirm similar findings in the STLF literature that there is little improvement gained by more sophisticated models compared with the persistence models ([13], [15]).

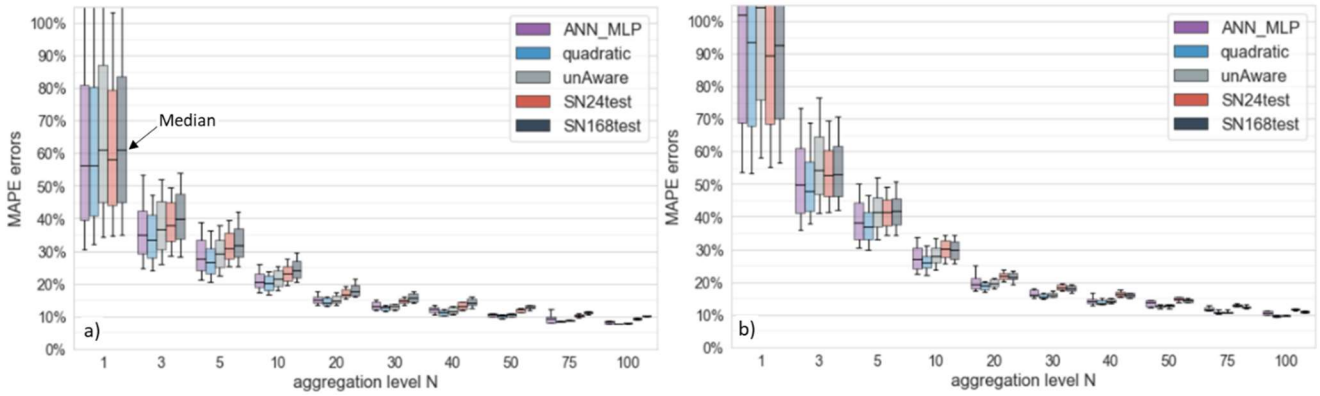

Figure 11 Errors for all Models by Household Aggregation Level (ANN Config2). Boxes and Whiskers as in Fig.10 a) LCL b) CER.

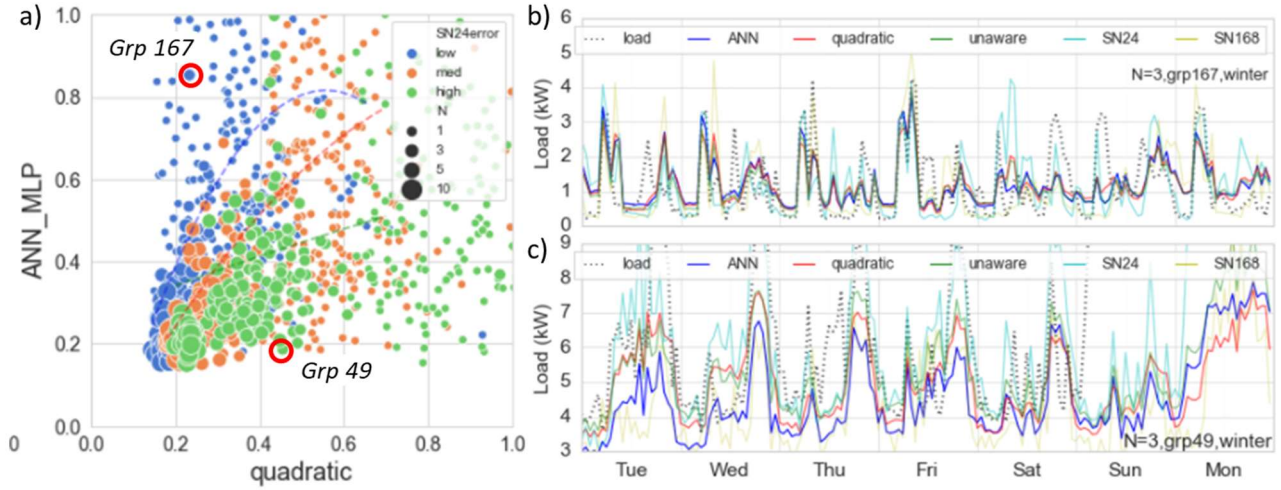

Figure 12 a) Distribution of MAPE Errors by Model Type for Aggregation N <=10. Classified by 24-hr Persistence Model Errors and Aggregation Level, N b) Load Profile and Model Forecasts for Grp 167 and c) Grp 49 for the Same Week (LCL).

3.4. Comparison of LCL v CER results

Forecasts generated from the Irish dataset have consistently larger errors. The weather inputs used for training may not be representative as they were taken from central Dublin whereas the smart meters were distributed throughout the country. The data includes a mix of rural and urban consumers with a wider diversity of load shapes. The persistence model results suggests equation (2) is a better average model than equation (1) for Irish customers. We have compared the distribution of daily household consumption and standard deviation across users for both datasets and found that the CER dataset has higher variance which may reflect the greater diversity of consumer types e.g. urban and rural.

3.5. Comparison with Similar Studies

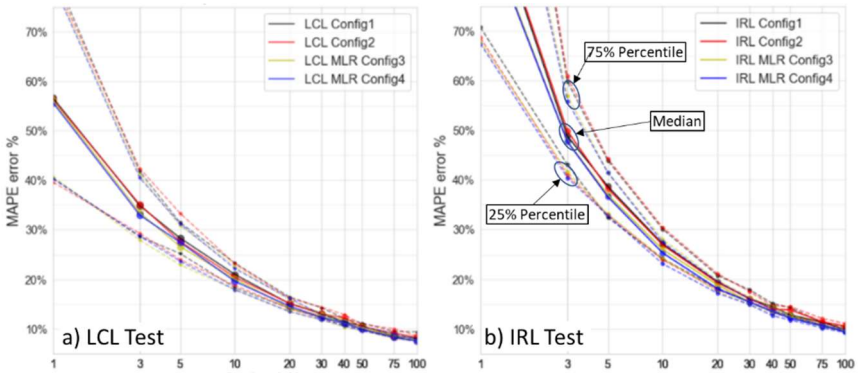

Figure 13 MAPE Errors (Median, 25% and 75% Quartiles) v Aggregation Level for ANN and MLR Model Configurations (Table 2) a) LCL b) CER

Similar studies which have looked at how errors scale with aggregation level for forecasts from hour to day ahead and have developed scaling factors according to data size and forecast horizon [19],[25] and [20]. We have adopted the scaling function from the most comprehensive such study [20] and applied to the ANN forecast results for both datasets. The scaling rule developed by the authors is given in Figure14 where W is the mean aggregate group load in Wh for each

Table 4 Optimized fit parameters for MAPE error scaling function.

	√α0	√α1
LCL	12.8617	0.0601
CER	21.2824	0.0220
R&S 2018	45.3000	1.4200

aggregation level N. We plotted the ANN 'config2' model error against the mean load for each group at all aggregation levels and fitted a similar function to the median MAPE errors (Figure14). We observe that the form of the function fits the error pattern well, but our coefficients are quite different (Table 4) to those from [20]. The average hourly consumption in both our data sets is ~ 0.5 kWh/hr whereas the dataset used in [20] is from California with an hourly average > 1 kWh. The forecast horizon is shorter for most of the models considered in [20]so we have used their published parameters for their SARMA day ahead aggregate forecast. These factors may account for some of the difference compared to our observations. We note that the errors increase more rapidly for the CER data at any given aggregation level. This may possibly be explained by the factors listed in 3.4

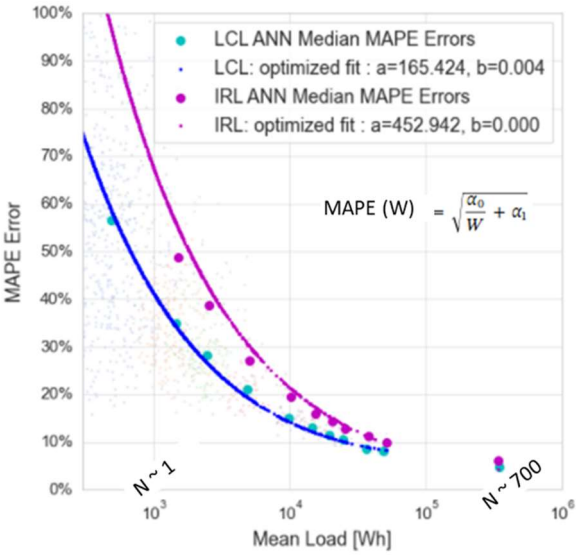

Figure14 Scaling of MAPE errors by Mean Hourly Consumption by Aggregation Group (Background Data are LCL MAPE Errors for ANN Model).

3.6. Practical Implications and Future Research

Since the error is highly dependent on aggregation size, it is desirable to cluster as many households as feasible for forecasting day-ahead demand. For example, an error of 20% for groups of 10 households may be acceptable for some applications such as DR. We suggest testing additional ANN models including recurrent or LSTM (long-short term memory) ANNs which can capture the temporal aspects of time series better than FFNN. Further optimization of the ANN parameters is recommended, for example, we used 2 hidden layers for training the CER ANN based on a marginally better performance over a single HL at the aggregate level. However, we did not have time to test if this was also better at smaller aggregations. A recent publication [22] has compared LSTM and FFNN results using a subset of the LCL data, finding that the LSTM model gave lower errors . It may be useful to classify consumers e.g. using a simple persistence model-based predictability measure or a clustering technique prior to running forecasts. We also recommend revisiting error metrics and repeating the above analysis with other metrics.

4. CONCLUSIONS

Key conclusions are a) Weather variables have minor explicit influence on forecast accuracy b) Load is accurately predicted to 5-6% error at 1 day ahead for the aggregate of ~ 700 households and there is significant predictability (< 20% error) for smaller group aggregates e.g. > 10 households. However, for individual users and very small clusters, all models show very large errors and most individual users are essentially unpredictable. c) The MLR models show better predictability than the persistence models for all aggregation levels d) ANN does not perform any better than MLR in our tests and sometimes gives results that are no better than the persistence models d) A small cohort of users are substantially more predictable than others and these tend to have low errors on all models provided their overall consumption is not anomalously low. e) There is a trade-off between the accuracy and the temporal resolution of the forecast f) Forecasts for Irish consumers have a much larger error compared with LCL equivalent. We conclude that forecasts for customers in a limited geographic setting are likely to be more accurate. We recommend a survey of end users to establish what is a useful error threshold for day-ahead forecasting.

5. ACKNOWLEDGEMENTS.

The author would like to thank Dr Edward Barbour for his support and supervision throughout the project. We also acknowledge CER Smart Metering Project -Electricity Customer Behaviour Trial, 2009-2010, accessed via the Irish Social Science Data Archive - www.ucd.ie/issd

6. REFERENCES.

[1] J. Ponoćko, *Data Analytics-Based Demand Profiling and Advanced Demand Side Management for Flexible Operation of Sustainable Power Networks*. Cham: Springer International Publishing, 2020.

[2] P. Goncalves Da Silva, D. Ilić, and S. Karnouskos, 'The Impact of Smart Grid Prosumer Grouping on Forecasting Accuracy and Its Benefits for Local Electricity Market Trading', *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 402–410, Jan. 2014, doi: 10.1109/TSG.2013.2278868.

[3] G. Gross and F. D. Galiana, 'Short-term load forecasting', *Proceedings of the IEEE*, vol. 75, no. 12, pp. 1558–1573, Dec. 1987, doi: 10.1109/PROC.1987.13927.

[4] E. Feinberg and D. Genethliou, 'Chapter12: Load Forecasting', in *Applied Mathematics for Restructured Electric Power Systems*, 2005, pp. 269–285.

[5] Hong, Tao, *Short Term Electric Load Forecasting*. Doctral Dissertation, N Carolina State University, 2010.

[6] T. Hong, P. Pinson, and S. Fan, 'Global Energy Forecasting Competition 2012', *International Journal of Forecasting*, vol. 30, no. 2, pp. 357–363, Apr. 2014, doi: 10.1016/j.ijforecast.2013.07.001.

[7] C. Kuster, Y. Rezgui, and M. Mourshed, 'Electrical load forecasting models: A critical systematic review', *Sustainable Cities and Society*, vol. 35, pp. 257–270, Nov. 2017, doi: 10.1016/j.scs.2017.08.009.

[8] T. Hong, 'Big Data Analytics: Making the Smart Grid Smarter [Guest Editorial]', *IEEE Power and Energy Mag.*, vol. 16, no. 3, pp. 12–16, May 2018, doi: 10.1109/MPE.2018.2801440.

[9] Y. Wang, Q. Chen, T. Hong, and C. Kang, 'Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges', *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, May 2019, doi: 10.1109/TSG.2018.2818167.

[10] J. Schofield, 'Dynamic time-of-use electricity pricing for residential demand response: design and analysis of the Low Carbon London smart-metering trial', May 2015, doi: https://doi.org/10.25560/25575.

[11] 'Comission for Energy Regulation (CER). (2012). CER Smart Metering Project - Electricity Customer Behaviour Trial 2009-10', Comission for Energy Regulation, 2012. [Online]. Available: www.ucd.ie/issda/CER-electricity.

[12] H. S. Hippert, C. E. Pedreira, and R. C. Souza, 'Neural networks for short-term load forecasting: a review and evaluation', *IEEE Transactions on Power Systems*, vol. 16, no. 1, pp. 44–55, Feb. 2001, doi: 10.1109/59.910780.

[13] E. Barbour and M. González, 'Enhancing household-level load forecasts using daily load profile clustering', in *Proceedings of the 5th Conference on Systems for Built Environments - BuildSys '18*, Shenzen, China, 2018, pp. 107–115, doi: 10.1145/3276774.3276793.

[14] N. Fumo and M. A. Rafe Biswas, 'Regression analysis for prediction of residential energy consumption', *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 332–343, Jul. 2015, doi: 10.1016/j.rser.2015.03.035.

[15] A. Veit, C. Goebel, R. Tidke, C. Doblander, and H.-A. Jacobsen, 'Household Electricity Demand Forecasting -- Benchmarking State-of-the-Art Methods', *arXiv:1404.0200 [cs, stat]*, Apr. 2014, Accessed: May 04, 2020. [Online]. Available: http://arxiv.org/abs/1404.0200.

[16] H. Shi, M. Xu, and R. Li, 'Deep Learning for Household Load Forecasting—A Novel Pooling Deep RNN', *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5271–5280, Sep. 2018, doi: 10.1109/TSG.2017.2686012.

[17] J. Moon, S. Park, S. Rho, and E. Hwang, 'A comparative analysis of artificial neural network architectures for building energy consumption forecasting', *International Journal of Distributed Sensor Networks*, vol. 15, no. 9, p. 1550147719877616, Sep. 2019, doi: 10.1177/1550147719877616.

[18] T. Hong and S. Fan, 'Probabilistic electric load forecasting: A tutorial review', *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, Jul. 2016, doi: 10.1016/j.ijforecast.2015.11.011.

[19] A. Marinescu, C. Harris, I. Dusparic, S. Clarke, and V. Cahill, 'Residential electrical demand forecasting in very small scale: An evaluation of forecasting methods', in *2013 2nd International Workshop on Software Engineering Challenges for the Smart Grid (SE4SG)*, May 2013, pp. 25–32, doi: 10.1109/SE4SG.2013.6596108.

[20] R. Sevlian and R. Rajagopal, 'A scaling law for short term load forecasting on varying levels of aggregation', *International Journal of Electrical Power & Energy Systems*, vol. 98, pp. 350–361, Jun. 2018, doi: 10.1016/j.ijepes.2017.10.032.

[21] Y. Wang, N. Zhang, Y. Tan, T. Hong, D. S. Kirschen, and C. Kang, 'Combining Probabilistic Load Forecasts', *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3664–3674, Jul. 2019, doi: 10.1109/TSG.2018.2833869.

[22] A. Mehdipour Pirbazari, M. Farmanbar, A. Chakravorty, and C. Rong, 'Short-Term Load Forecasting Using Smart Meter Data: A Generalization Analysis', *Processes*, vol. 8, no. 4, p. 484, Apr. 2020, doi: 10.3390/pr8040484.

[23] 'Acorn-User-guide'. https://acorn.caci.co.uk/ (accessed Apr. 16, 2020).

[24] R. E. Edwards, J. New, and L. E. Parker, 'Predicting future hourly residential electrical consumption: A machine learning case study', *Energy and Buildings*, vol. 49, pp. 591–603, Jun. 2012, doi: 10.1016/j.enbuild.2012.03.010.

[25] S. Humeau, T. K. Wijaya, M. Vasirani, and K. Aberer, 'Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households', in *2013 Sustainable Internet and ICT for Sustainability (SustainIT)*, Oct. 2013, pp. 1–6, doi: 10.1109/SustainIT.2013.6685208.

[26] S. Haben, J. Ward, D. Vukadinovic Greetham, C. Singleton, and P. Grindrod, 'A new error measure for forecasts of household-level, high resolution electrical energy consumption', *International Journal of Forecasting*, vol. 30, no. 2, pp. 246–256, Apr. 2014, doi: 10.1016/j.ijforecast.2013.08.002.