Household Classification using Smart Meter Data

Paula Carroll[1], Tadhg Murphy[1], Michael Hanley[1], Daniel Dempsey[1] and John Dunne[2]

[1] Centre for Business Analytics, School of Business, University College Dublin, Ireland

[2] Central Statistics Office, Cork, Ireland

Abstract: This paper describes a project conducted in conjunction with the Central Statistics Office of Ireland in response to a planned national rollout of smart electricity metering in Ireland. We investigate how this new data source might be used for the purpose of official statistics production. This study specifically looks at the question of determining household composition from electricity smart meter data using both Neural Networks (a supervised machine learning approach) and Elastic Net Logistic regression. An overview of both classification techniques is given. Results for both approaches are presented with analysis. We find that the smart meter data alone is limited in its capability to distinguish between household categories but that it does provide some useful insights.

Keywords: Neural Network; Elastic Net Logistic Regression, Classification system; Household composition; Smart Meter Data

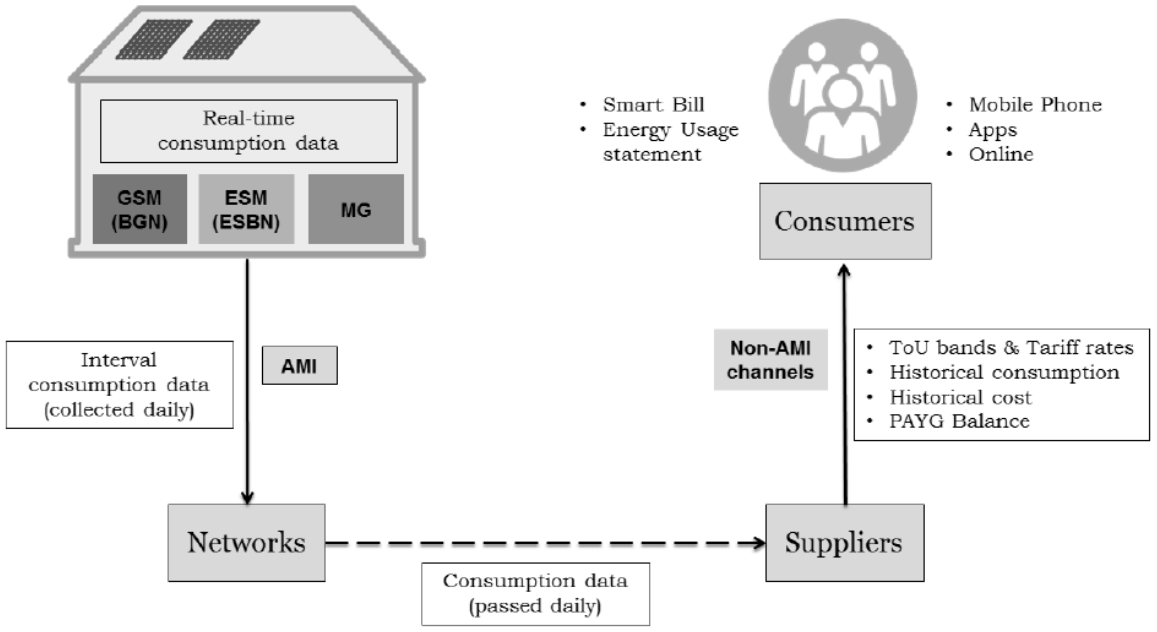
1. *Introduction*

Smart Meters (SM) in the residential sector are seen as a key factor in the success of EU targets for reduction in greenhouse gases and increases in the use of renewable energy (Europea, 2014). An SM system has an electronic meter which sends electricity load data to and receives price data from the service provider. The Irish Commission for Energy Regulation (CER) initiated the National Smart Metering Programme (NSMP) in 2007 and Customer Behaviour Trials (CBTs) took place during 2009 and 2010 to assess the performance of SMs and their impact on consumer behavior. The purpose of the CBTs was to gauge customer response to price incentives. The anonymised data gathered during the trial are available for research purposes (ISSDA, 2014).

It is anticipated that a full rollout of SMs in Ireland will commence in 2019. Each consumer will have an individual meter to enable each residential household better manage its electricity usage. More recently, CER have announced the high level design decisions for the NSMP. Figure 1 shows an overview of the proposed architecture, CER (2014). We see that consumption and price data will be exchanged but no information about the building, appliances or residents will be provided.

Smart Meter Data (SMD) opens up new opportunities for researchers, businesses and public sector organisations. In particular, the potential role of SM data in the production of official statistics is of interest to national statistical institutes and is the focus of this paper. New data sources such as SMD have the potential to provide valuable information and insights about not only energy consumption but also household consumption and possibly, the subject of this study; household composition.

Like most countries, the Central Statistics Office of Ireland (CSO) is exploring ways to modernise how it calculates population estimates, (Dunne, 2015). The focus of this research is an exploration of SMD to estimate household composition. Household composition is a classification of households by size and relationship type between the household members. It is currently established in Ireland in a costly census every five years. This involves the distribution of census forms to every household in the state and the subsequent collection of these forms. The cost of the 2011 census was €55 M. The SMD gathered during the CBT trial is used as to attempt to answer our research question.



GSM – Gas Smart Meter provided by Bord Gais Networks (BGN)
ESM – Electricity Smart Meter provided by Electricity Supply Board Networks (ESBN)
MG – Micro-Generation meter provided by parties yet to be determined
AMI – Automated Meter Infrastructure
ToU – Time of Use
PAYG – Pay As You Go (enhanced form of PrePayment)

Figure 1: NSMP High Level Design CER (2014)

Research question: Can household composition be estimated from analysis of SM electricity usage?

We evaluate two techniques to classify households; Neural Networks and Elastic Net Logistic Regression. While existing CSO household composition categories cannot be readily identified, useful insights can be gained from SMD analysis. In particular, the models are useful in identifying households of single persons. The model performance worsens as the number of persons in a household increases.

The remainder of this paper is structured as follows: Section 2 outlines the challenges and opportunities for national statistics institutes in new data sources such as SMD; Section 3 describes the classification methods and data issues. Section 4 gives results and analysis; Sections 5 and 6 include a discussion and conclusions of the work.

2. Challenges and opportunities for National Statistical Institutes (NSIs)

The functions of the CSO are spread across many areas with responsibility for the collection, compilation, extraction and dissemination for statistical purposes of information relating to economic, social and general activities and conditions in Ireland. Like most NSIs, the CSO is exploring ways to modernise how it operates and are trying to increase and improve the services they offer despite the growing costs of data collection and processing, and ever more challenging fiscal environments. A survey of the evolving National Data Infrastructure in Ireland is given in Dunne (2015). A strategy which focuses on efficient public administration rather than purely the production of official statistics is envisaged. This may be accomplished through the linking of administrative data registers covering persons, business and property. Currently, projections of the population on an annual basis up to 2046 are based on projection forward from the 2011 census base under a chosen set of assumptions governing births, deaths and net migration. Dunne (2015) describes some emerging opportunities for

future censuses that may exploit administrative data registers either in conjunction with or as a substitute for primary data collection.

Seyb et al (2013) describe the strategy implemented by Statistics New Zealand to improve and standardise processes in official statistics production. One goal in their change programme is to maximise the use of administrative data as a source wherever possible, with surveys filling gaps in information needs. This is a reversal of traditional survey-based data gathering strategies. Seyb et al (2013) describe how value can be extracted from a specific administrative data source where the data is well formatted and well defined. They give an example where tax data reference numbers used by Inland Revenue agencies are already mapped to business registers, so matching and coverage issues are easy to resolve. The data items in that instance are well defined financial variables.

Other administrative or new data sources may not be as amenable to adaptation for NSI purposes. The focus of our work is on SMD as a potential new data source for the CSO. Every household uses electricity but the data derives from electricity markets and was not intended for NSI usage. In this paper we outline the first steps toward harvesting value from SMD data.

2.1. Evaluating SMD data for Official Statics Production

The Irish CBT SMD has been explored to identify factors influencing domestic energy consumption. Dwelling characteristics (such as dwelling type, age and electrical appliances) and occupant characteristics (such as household income, age of household members, household composition) have been used to explain energy consumption. See for example McLoughlin et al (2012). The reverse, using consumption to predict occupant characteristics, has received little attention (Newing, 2015). It should be noted that dwelling and socioeconomic information about the CBT participants were used by McLoughlin et al (2012). Such information was available in the CBT but will not be available through the smart meter itself, (Van Gerwen et al, 2006 and CER, 2014).

2.2. *The CBT data*

As noted, SMD data derives from electricity markets and the focus of the CBT trial was consumer responsiveness to pricing structures. However, the CBT data gives an indication of what SMD looks like, its volume and velocity and allows us to attempt to answer our research question.

The CER used a stratified random sampling framework to invite consumers to participate in the CBT. This ensured the sample was broadly representative of the population in terms of household size and other socio-economic indicators. Over 5,000 consumers were initially recruited, further details are given in Section 4.1. Each consumer represents a household, i.e. a number of persons sharing a single residential unit.

An incentive of €25 for completing a pre-and post-trial survey was offered. An additional incentive for participation was the possibility of lower electricity bills during the trial depending on the consumer's response to the pricing schemes. The surveys were conducted by computer assisted telephone interviewing and focused on participants' views on attitudes to electricity usage and expectations of the trial, the dwelling and electrical appliances. Questions on demographics and social relationships between household members were limited as they were not the focus of the CBT study.

The CBT recorded a meter reading of the electricity usage of participating consumers at half hourly intervals over the duration of the trial. Each household meter produced 269 MB of such time series usage data during the trial. There are over 1.6 M households in Ireland. This gives an indication of the type and volume of data associated with electricity consumption per household that will be available after national SM rollout.

The volume of such data presents a significant challenge for NSIs such as the CSO which does not have a history of dealing with high volume data other than its own primary (well structured) sources. The infrastructure required to deal with such data volumes has not been investigated in this study.

This study focuses instead on a data processing pipeline and analytics techniques to produce meaningful insights on household composition from a SMD data stream.

3. Classification Techniques

The goal of classification in this paper is to assign a household composition category to a household based on its SM electricity usage. The parsimony principle tells us that classification models with a small number of explanatory variables (EVs) are preferable. In this paper, the dependent variable (DV) is the household classification and the EVs are drawn from the SMD data. Further EVs relating to participants' dwelling type and the type of electrical appliances used, are available in the CBT surveys. However such information will not be available with the SMD after rollout, only the electricity usage data will be available. So, only EVs from the SMD data are used in this proof of concept study.

We use the CBT survey response on household composition to label the SMD. The labelled SMD data is processed through a data reduction pipeline to yield a set of EVs suitable for model building. The data reduction process is described in Section 4. This labelled data allows us to use a supervised machine learning approach. We train and test a Neural Network to identify household composition based on SMD usage. These results are compared with those from a statistical model, namely Elastic Net Logistic Regression.

3.1. Regression

Regression is often used as a benchmark for classification tasks. Multiple linear regression models the linear relationship between DVs *y* and a set of EVs *x*. The general form is $y = \beta_0 + \sum \beta x$ where the coefficients β are calculated so as to minimise some loss function, such as the sum of squared error $\|y - \beta x\|^2$.

A regularisation term may be added to the loss minimisation objective function to achieve parsimony and reduce overfitting. Two popular approaches to regularisation are Ridge regression (Hoerl and Kennard, 1988) and LASSO (Tibshirani, 1996). Ridge regression adds a squared 2-norm penalty term on the coefficients, used to reduce the variance inflation due to correlations in the explanatory variables. Least Absolute Shrinkage and Selection Operator (LASSO) adds a 1-Norm penalty term which has the effect of shrinking coefficients, possibly all the way to zero, thus performing what can be considered a continuous variable selection as opposed to discretely dropping variables outright. Elastic Net harnesses both Ridge and LASSO regularisation approaches by taking a linear combination of both norm penalties (Zou and Hastie, 2005). The Elastic Net is fit by minimising $\|y - \beta x\|^2 + \lambda[\![\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2]\!]$.

The $\|\beta\|_1$ term is the LASSO penalty. The $\|\beta\|_2^2$ is the ridge penalty. The λ parameter is nonnegative and controls the 'strength' of the regularisation. A larger value of λ corresponds to greater variance reduction in the coefficient estimates but induces stronger bias. A value of λ = 0 corresponds to standard least squares regression. The α parameter takes values between 0 and 1 and controls the weight of the penalties. An α > 0.5 puts more weight on the variable selection properties of the LASSO, while α < 0.5 puts more weight on the correlation regularisation properties of the ridge.

Since linear regression models are linear by their nature, they are not well suited where the relationship between the inputs and outputs is not well defined or linear as is the case for electricity consumption and household composition. Generalised Linear Models (GLM) such as logistic regression can be used to overcome this limitation and to attempt to improve the model fit. GLMs extend the ideas behind linear regression. The dependant variables arise from the exponential family and are related to the EVs by a link function *f*, $f(E[y]) = \beta_0 + \sum \beta x$. The logit function can be used as the link function to predict categorical variables in a logistic regression model. This allows binary and multinomial

classification, where logit(x) = $\ln\left(\frac{x}{1-x}\right)$ is the log odds. This forces the output to be a value between 0 and 1 which can be interpreted as a probability that the outcome belongs in a certain class.

The regression coefficients are usually estimated using maximum likelihood estimation. Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximise the likelihood function, so an iterative process such as Newton's method is used instead.

The Elastic Net approach can also be used to reduce overfitting of the logistic regression model.

3.2. Machine Learning and Neural Networks

Standard statistical techniques are based on assumptions that the data items have been sampled independently and from the same distribution. Machine learning (ML) offers several techniques for intelligent data analysis and knowledge generation through generalisation where such assumptions may not hold (Hand, 1996). Techniques include: Association rules, decision trees, inductive logic programming, Support vector machines, Clustering, Bayesian networks and (artificial) neural networks (A)NNs. NNs are discussed in detail here, a review of the other techniques is beyond the scope of this paper. NNs are often preferred when large noisy training samples are available and the relationships may be non-linear. Disadvantages of NNs include their "black box" nature and the empirical nature of the model development.

The aim of a ML model is similar to that of a regression model. It aims to model how the set of inputs (called *features* in ML parlance) relate to the set of outputs. However the approach to creating and fitting the model differs from regression. The ML model is learned from a training data set. In supervised learning, the outputs for the training data are known (labelled). A ML learning algorithm adapts the model in response to the training data to improve the fitting of the input/output relationship.

Perhaps the most important concept in ML is that of generalisation. The algorithm should produce sensible outputs for inputs that were not encountered during learning Marsland (2009). Over-fitting occurs when a model fits only the training data, meaning that it is not a general function approximation. It has instead begun to learn the noise associated with that specific training data set. To ensure that over-fitting does not occur, the data is usually split 60:20:20 into training, validation and testing sets. The learned system is evaluated on the validation data set to assess the ML model fit before being used on unseen test data.

ML can be used to perform classification. We assign the input(s) to discrete output categories. Testing is performed to evaluate the model in terms of the performance of its classification when it is given new data without class labels. The actual class labels of each input are compared with those assigned by the algorithm. *Accuracy* is defined as the percentage of correct matches, that is, the number of correct (or true) household classifications in our case divided by the total number of tests:

$$Accuracy = \frac{Number\ correct\ classifications}{Number\ of\ tests}.$$

The accuracy metric can be misleading when the data are dominated by a high number of entries from a single class. This issue is explored in more detail in Section 3.3. A classifier that simply predicts the dominant class will have high accuracy but could not be regarded as a good classifier. Other commonly used error metrics are discussed in more detail in Section 4.3.

NNs are ML algorithms modelled on the function and topology of the human brain. NNs have successful applications in diverse areas from credit card fraud detection (Patidar,2011) to forestry management (Hickey et al, 2015) to energy consumption modelling (Aydinalp, 2002). Aydinalp (2002) favoured NNs over statistical models due to the simplicity of NN development and the accuracy of the estimate. They found that NNs were capable of modelling nonlinear electricity consumption relationships outperforming statistical approaches.

In the brain, nerve cells called *neurons* function as simple processing devices. Neurons can be described as simple mathematical functions. A general form for a single neuron is $y = g(w_0 + \sum w\,x)$ where g is called the transfer function, often a sigmoid or hyperbolic tangent function. The w are weights which are analogous to the coefficients in a regression model and w_0, known as the bias, is analogous to the regression intercept coefficient.

Figure 2 shows a representation of a simple NN with EVs x_i, a single neuron and a single output. NNs consist of an array of neurons that form a connected network (Hopfield, 1984, Zhang and Zhang, 1999). A Multi-Layer Perceptron (MLP) is a feed-forward NN consisting of an input layer, one or more hidden layers and an output layer.

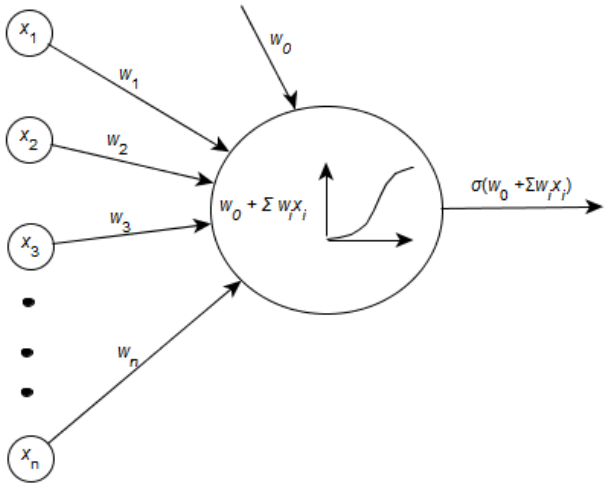


Figure 2: Model of a simple NN

An iterative approach called the error back-propagation (BP) algorithm is used in a MLP to estimate the weights during the training stage. BP consists of two passes through the network: a forward pass and a backward pass. The weights *w* are fixed in the forward pass. An input vector from the training data propagates through the entire network to produce a set of outputs. The difference between the produced output and target value is calculated as the error. The error is similar to the loss function in a regression model. On the backward pass, the weights *w* are adjusted according to some error-correction rule (such as a gradient descent function) to reduce the error. In this way, the NN response is moved closer to the desired output. Termination criteria for the iterative model fitting are used to stop the BP algorithm when improvements fall below a threshold. The NN weights are then finalised and the NN is ready for the test phase.

The empirical approach to NN model development means there is no guarantee that the final NN weights are the global optimal weights. They may reflect local optima. Nor is there a guarantee that the selected topology of hidden layers is optimal.

3.3. *Data Issues*

In this section we identify some of the issues that arise for NSIs interested in exploring SM data. When dealing with large volumes of data, analysts have to decide on a data reduction scheme which adequately represents the data. The reduced data representation conveys most of the information while being easier to store and facilitates ease of computation. The choices are to aggregate the data (using representative measures), use samples of the data or to apply more advanced data reduction algorithms.

Recall that the EVs used as inputs to NN models are called *features*. As the number of dimensions (or possible EVs) in the search space increases, the amount of data needed to provide the algorithm with sufficient training examples increases rapidly. This can lead to an explosion in the amount of training data required as well as lengthening the training time for the NN. This issue is known as "the curse of dimensionality". Dimension reduction techniques can be used to address this issue, see for example Han (2006).

Characterisation of Time Series (TS) data such as SMD is discussed in detail in Liao (2005) and Wang (2006). The lower level half-hour granularity per meter gives a better picture of what is happening in each household type than would be apparent by looking at aggregated SMD daily totals. The individual consumer's load profile may offer a unique fingerprint to aid classification. However, it is not desirable to work directly with raw data that are highly noisy. Instead application dependant extracted features are used. In addition, a choice on the length of the TS is required. In general, larger sample sizes yield better population estimates with lower variability. However, in the case of TS data, longer series may actually increase variability due to any underlying trend. The choice of the TS window length is one of the many design parameters and is an open ML research question. Varying length sequences can be empirically evaluated and/or adaptive windowing (similar to the lag methods in ARIMA models) can be used to weight the contribution of varying length sub-sequences within the TS data.

In many applications the training data is unbalanced, that is, some categories are under or over represented. It is important to distinguish between imbalance in the training data sets and representativeness of the population of the sample. For example, in the classification of defective products at the end of an assembly line, the majority of products, perhaps 90%, are good since they meet the required standard. The remainder fail and are deemed defective. While the imbalance reflects the distribution of the items in the population, traditional feed forward NNs have difficulty learning from unbalanced datasets. NNs need to see an equal number of defective and good products during the training phase to learn how to distinguish them. Otherwise the NN prioritises the class seen in the majority of samples and treats the minority class as noise (Murphey, 2004). In the production example, the NN would classify all products as good and 90% of the time would be correct. This results in misleadingly high accuracy values for the model. Several error metrics are used to interpret the results in conjunction with the accuracy measure. In addition, resampling or over-sampling can be used to address issues of unbalanced training data.

Lastly, concerns about the privacy of the individual arise with SMD (Molina-Markham, 2010, McKenna, 2012). These include that the SM signals may be intercepted for illegal purposes by third parties and that SMs allows surveillance of the individuals' usage rather than simply tracking usage for billing purposes. The CSO adheres to the UN Fundamental Principles of Official Statistics and seeks to balance the public interest with concerns for privacy of the individual. The CSO was established statutorily under the Statistics Act, 1993 (Statistics, 1993). This act includes articles on statistical confidentiality so the CSO is well positioned to explore new data sources such as the SMD.

4. *Material and Methods*

In this paper, consumers are classified and assigned to a household category based on their electricity usage. NNs are selected to perform the ML household classification due to their ability to work with high volume noisy data and learn non-linear relationships. Elastic net logistic regression is selected as a comparative GLM statistical approach. We reduce the individual consumer TS streams to sets of possible features (explanatory variables) and select the most useful subset of features. We evaluate the model performance over varying TS window lengths and compare results from both unbalanced and balanced training data sets.

4.1. Data Pipeline

Some information about the age of household members is available from the CBT pre-trial survey. A limiting factor of the CBT survey from the household classification perspective is that detailed age information is given for the Head of Household only. The remaining members of the household are classified as either *under 15 years of age* or *15 years of age and older*. In addition, no information is given on family unit group, for example whether the household consists of a married or cohabiting couple or single parent with children etc. Existing CSO household composition categories distinguish family types, for example *cohabiting couple with children* or *husband and wife with children*. This categorisation is useful in social analysis and understanding changing demographics but we would not anticipate a difference in electricity usage of based on marital status. Indeed, in a sign of changing times, the marriage equality referendum passed in Ireland in 2015, may see the need for the development of new household categories such as *husband and husband with children*.

For the purpose of this smart meter study, an alternative simple household categorisation system was developed according to the numbers of adults and children as shown in Table 1. These 16 household categories were chosen as they match 95% of the existing CSO categories and represent the majority of the CBT data. An even simpler classification based on the number of persons per household was also considered.

Table 1 shows how representative the CBT sample is of the 1.6 M households in Ireland. The final two columns of Table 1 show the similarity of the CBT household distribution to the percentage of households by number of persons according to the 2011 census (CSO, 2011). The minor gap is households with eight or more persons as none participated in the CBT, this category accounts for 5% of all households in Ireland.

Table 1: Household category description

Category	Adults	Children	Meter Count	Post-processing Count	Num persons	CBT Distribution	CSO Distribution
A	3	2	41	39	5	1%	2%
B	3	1	106	105	4	3%	3%
C	3	0	450	440	3	11%	10%
D	2	5	9	9	7	0%	0%
E	2	4	49	48	6	1%	1%
F	2	3	158	147	5	4%	4%
G	2	2	338	331	4	9%	8%
H	2	1	246	244	3	6%	7%
I	2	0	1,264	1,251	2	32%	27%
J	1	1	59	59	2	2%	2%
K	1	0	726	718	1	19%	24%
L	4	1	64	64	5	2%	2%
M	4	0	289	283	4	7%	5%
N	5	1	20	20	6	1%	1%
O	5	0	92	92	5	2%	1%
P	6	0	20	20	6	1%	0%
≥ 8			0	0	≥ 8	0%	5%
Total			3,931	3,870		100%	100%

A significant work component of this study was to convert the CBT SMD data to household classifications. Data pre-processing absorbed approximately 65% of the project man hours. Over 150 Million data points of usage are included in the SMD trial data in multiple CSV files. Each SMD usage file consists of 3 columns corresponding to a unique household Meter ID, timestamp and electricity consumed during 30 minute intervals in kWh. In order to allow a valid comparison, SMD from the six month benchmark period from July to December 2009 were considered. Price incentives were evaluated during the later months of the CBT trial. Some work on consumer behaviour and their responsiveness to tariff changes is described in Di Cosmo (2012). Such work could be used to estimate the likely changes in household electricity usage patterns in response to tariff changes.

The data were prepared using the open-source package R (R, 2013). Standard workplace laptops with 8 GB RAM were used for light data pre-processing tasks. Data for households who had not completed the survey were removed leaving 3,931 sets. Meters with missing data were also removed. The data reduction and model building was then carried out using R on the Stokes supercomputer with 7,680 GB of RAM at the Irish Centre for High-End Computing.

Bousquet (2002) discuss the use of sensitivity analysis to evaluate changes in ML algorithm outcomes to changes in the training set. We were particularly interested in assessing the impact on the model performances of the window length of time series used as the training data. Five different time series window lengths ranging from one day to six months were chosen so that the sensitivity of the classifiers could be empirically assessed.

Feature values for the five different time series windows were calculated on the Stokes supercomputer in the data reduction step. The features are the EVs or inputs for the classification models, further details are given in Section 4.2. The prepared data per meter was then labelled with a household classification category. These five files containing the feature values for the five different time series windows were then ready for use in creating and testing the classification models.

4.2. Data Reduction and Feature selection

Table 2 shows a summary of the extracted features. Some are suggested in McLoughlin et al (2012). Others are standard descriptive statistical measures typically used in NN time series modelling (Wang, 2006). The remaining features were identified from analysis of the diurnal usage patterns of individual household categories to spot distinctive features which may be unique to a household category. One such example is “morning peak”. It was noted that households with children generally had a more pronounced morning peak.

Twenty one features were calculated for each meter to summarise each household’s unique load profile over the five time series windows. Detailed descriptions are included in Appendix 1. The raw numeric input SMD data were standardised to between -1 and 1. Standardising is carried out to bring all variables into proportion with one another. Features that demonstrated multicollinearity with high inter-correlation coefficients were removed, leaving 18 input features or explanatory variables.

Outlier analysis was performed on the summarised data to remove any outlying households that might disrupt the performance of the classifiers, leaving 3,870 meters. Individual data within the meters was not subjected to any outlier analysis, instead this was performed on the aggregated data for each meter. This approach ensured that potentially useful data within individual meters was not removed but that outlying households were removed before the data was input to the classifier. For example, increased usage on a cold day was not deemed outlying. The local outlier factor algorithm which is a density-based outlier detection approach was chosen for this task. It can be computationally expensive as the approach involves the calculation of *k*-nearest neighbours. Breunig et al (2000) argue that this approach is more subtle than a simple binary outlier classification and allows the degree of closeness within a neighbourhood to be accounted for.

Table 2: Model inputs*Indicates a feature that was not selected.

Index	Feature (EV)	Short Description
1	Mean*	Mean energy consumption
2	Max	Maximum energy consumption
3	ToU Max	Time of day at which maximum consumption occurs
4	TEC	Total energy consumption
5	MDM	Mean daily maximum energy consumption
6	Load Factor	Ratio of daily mean to daily maximum energy consumption
7	Variance	How far the energy consumption is spread out
8	SD	Standard deviation from the mean
9	Range	Difference between highest and lowest energy
10	Interquartile range (IQR)	Measure of spread of middle half of data
11	Morning Max	Maximum energy use in the morning
12	Morning Peak	Height of the morning peak energy consumption
13	Morning Range	Morning maximum minus minimum before 10am
14	Weekday Area	Area under the curve for weekday consumption
15	Weekday Midpoint*	Area under the curve for weekday consumption divided by 2
16	Weekday Centroid	Time of day at weekday midpoint
17	Weekday AM Slope	Slope of the Morning Peak
18	Weekend Area	Area under the curve for weekend consumption
19	Weekend Midpoint*	Area under the curve for weekend consumption divided by 2
20	Weekend Centroid	Time of day at weekend midpoint
21	Weekend AM Slope	Slope of the morning Peak

4.3. Model Development

Two classifier approaches were evaluated. The first was a binomial classifier asking a binary question; whether a particular meter belonged to a particular household category. Classifier output greater than 0.5 was labelled as true (yes). Classifier output less than 0.5 was labelled as false (no). The advantage of a binomial approach is that only a single output is required. It was expected that the classifier would be better able to partition the dataset. The disadvantage was that the model had to be run separately for each household category and so involved extra data manipulation.

The second approach was a multinomial classifier asking which household category a meter belonged to. The output produced by the classifier is a vector of values between zero and one. These vector components are interpreted as probabilities that the meter belongs to the household categories. The household category with the highest probability is the most likely category to which the meter belongs. The advantage of the multinomial approach is that only one model is required and less manipulation of the data is needed. However, as the multinomial classifier has multiple outputs, it could potentially lead to a reduction in accuracy. Lower accuracy was anticipated as some overlap of electricity usage between classes was expected.

The "glmnet" package in R was used to implement the elastic net logistic regression models (Friedman et al, 2009). For all models α was set to 0.25. This puts more weight on the ridge penalty which averages correlated groups but still allows for some feature selection. 10-fold cross validation was used to set λ based on the misclassification error rate. For each Elastic Net model, 70% of the data was used as a training set, the remaining 30% was used for testing the predictive power of the

models. The "caret" package in R was used for splitting the data into training and test sets (Kuhn et al, 2014).

The R "nnet" NN package was used to build a single-hidden-layer NN by selecting the number of units in the hidden layer, the initial random weight, and the weight decay (Ripley and Venables, 2011). The "neuralnet" package was also used as it allows a choice of training algorithms and the number of hidden layers (Fritsch et al, 2016). Training of the NNs was carried out by back propagation, resilient back propagation with backtracking, resilient back propagation without backtracking and a modified globally convergent approach. The input data for the NNs was split into three subsets in ratios 60:20:20 for training, validation and testing. The training data set was sampled at random without replacement. From the remaining dataset, 50% were sampled at random without replacement to create the validation set with the remaining meters forming the test set.

The performance and suitability of all models were assessed under the headings of accuracy, Sum of Squared Error (SSE), Root Mean Squared Error (RMSE), Sum of Cross Entropy (SCE), Coefficient of Variation and Pseudo R^2. Confusion matrices of actual (row) versus predicted (column) values in each class were also produced. A good classifier exhibits a diagonally dominant matrix. Further details of the error metrics are available in Appendix 2.

For the binomial NN models, the RMSE was computed at each iteration of the NN development for both the training set and the validation set, and the SCE was used for the multinomial model. The training of the network was stopped when the RMSE/SCE error using the validation set registered two consecutive increases. A value of two was chosen as stopping after one increase might be premature and the increase might only a once-off result in a general trend of decreasing error. More than two consecutive increases was categorised as a trend of increasing error. This check found the point at which the training algorithm had started to over-fit the data.

The sensitivity of the binomial and multinomial NN models on both unbalanced data and balanced training data and on the five TS windows described in Section 4.1 were evaluated. Computational results are presented in Section 5.

4.3.1. Unbalanced Training Data

As noted in Section 3.3, a balanced number of training samples is preferred for ML classification so that one category does not bias the prediction output. Table 1 highlights the imbalance in the CBT which is a concern for training the NN. The number of sample households consisting of two adults and no children (1,264) exceeds any other household type in the trial. It is not a concern for the representativeness of the CBT data.

We used stratified sampling to build the training set for the Elastic Net unbalanced models. We sampled 70% of each category instead of simply taking 70% of the entire data. The value of λ in the elastic net was chosen via 10-fold cross validation where the validation error is the misclassification rate. Separately, the data was split 60:20:20 into training, validation and testing sets for the NN. The models were then applied to the test sets and a full set of error metrics was calculated.

4.3.2. Balanced Training Data

Under-sampling (He and Gracia, 2009) was used in order to achieve balance in the training data. This technique, for the binomial models, is to only sample enough records from the majority class so that it equals the number of records in the minority class. This is more suited to situations where large datasets are being analysed as it has the advantage of reducing the training time by effectively reducing the size of the training set. Recall the 16 different household composition categories described in Table 1. It shows that following the pre-processing step, the number of households in each of these categories ranged from 9 to 1,251 with a total dataset size of 3,870. For the household category with 9 meters this meant that the number of entries in the *true* class was 9 and the number of

entries in the *false* class was 3,861. To perform under-sampling on this category required sampling 9 records from the majority *false* class of 3,861, meaning that the size of the dataset for classification for this category was 18. This under sampling was repeated across each of the household categories to allow classification on balanced data.

For the multinomial model, an equal number of meters in each household category were sampled. As the minimum sample size was 9 this required sampling 9 from each of the remaining categories and training the classifier with 9 instances of each category. This is too small for ML algorithms, so it was decided to only analyse categories where the number of records in the category was greater than 100 meters. This choice ensured that a minimum sample size of 20 was achieved for a 60-20-20 cross validation split when developing the NN model. Eight of the 16 categories met this selection criteria, namely B, C, F, G, H, I, K and M. These categories accounted for 3,519 (91%) of the CBT meters after pre-processing and is representative of 86% of the population of 1.6M households in Ireland.

5. ***Results***

Figures 3 and 4 show examples of the weekday and weekend daily usage averaged over a six month period for household categories C and H. Category C is a household with three adults. Category H consists of two adults and one person aged under 15. Weekdays are shown as a solid line, weekends as dashed. These are typical of the diurnal usage pattern showing a peak corresponding to the start of the day, some activity at lunch time and a peak corresponding to preparation of an evening meal.

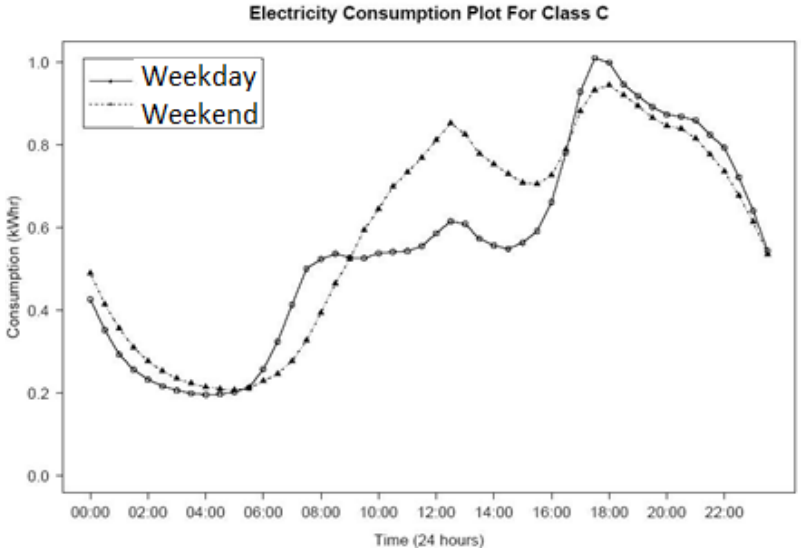


Figure 3: Household category C (three adults)

A box plot of the mean daily usage in Figure 5 highlights the increasing trend in mean values as the number of occupants within the house increases. It was expected that an increase in mean consumption would allow the classifiers to better distinguish between household categories. We also see the variety in the degree of dispersion and shape of the distribution across the household categories. The use of the local outlier factor algorithm means that only households that are relatively extreme were removed during pre-processing. Recall also that variance, standard deviation and IRQ are among the extracted features in Table 2.

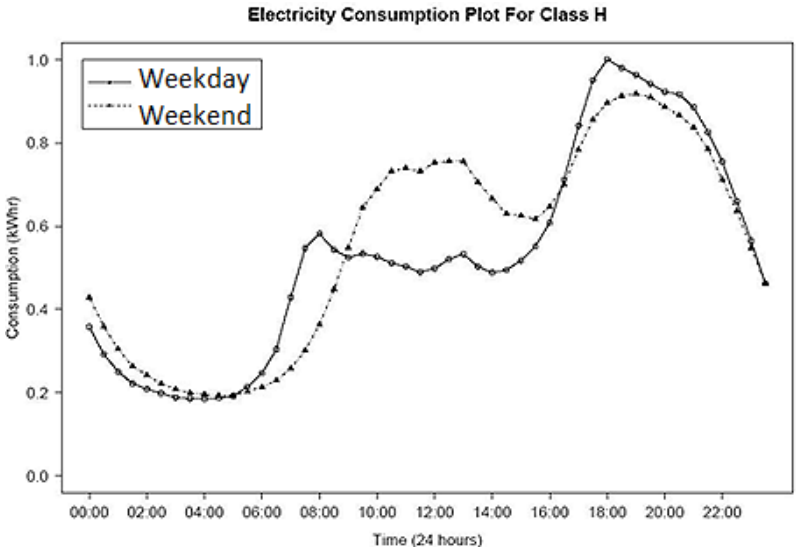


Figure 4: Household category H (two adults plus one child)

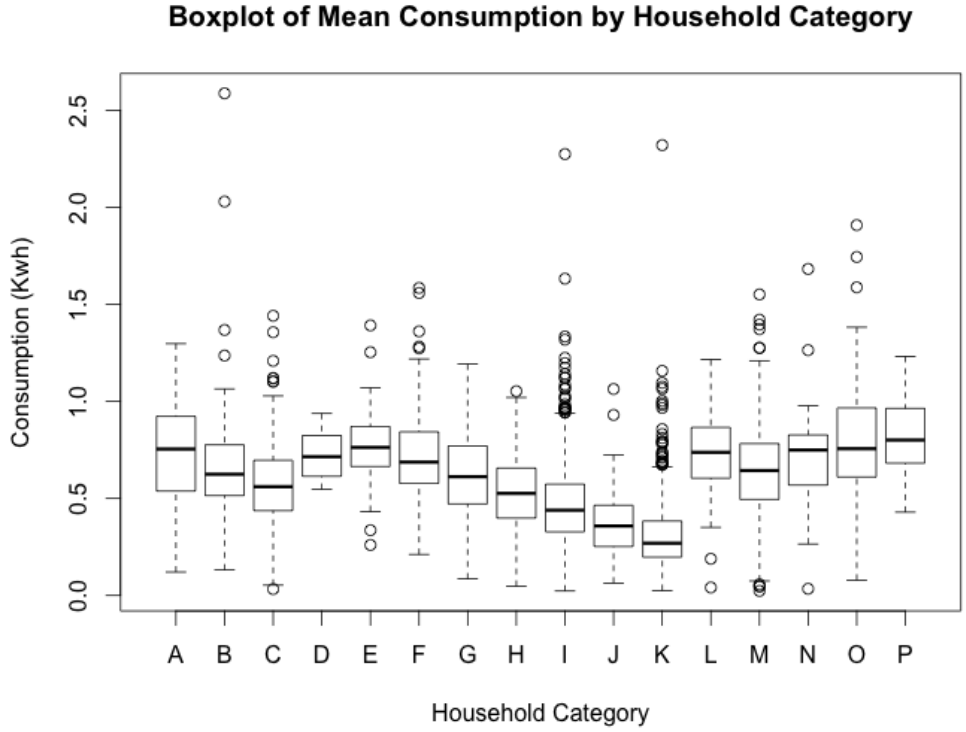


Figure 5: Boxplot of mean daily usage per household category

5.1. Classifier Results

The results from the Elastic Net Logistic regression model (EN) and from the Neural Net (NN) models were quite similar, see Table 3. Results for the simpler classification scheme (of numbers of persons) were not significantly better. In some cases it was slightly better, in other cases, it was slightly worse. In the interests of brevity, we present the results for our number of adults/children classification scheme (which is detailed in Table 1).

The six month TS window produced the best performance. The performance benefits in the longer time frame varied in comparison to other time windows. The six month window may capture some long distance interactions or seasonality. Again, in the interests of brevity we present the results for the six month window only.

Balanced training data gives the better results for both the EN and NN binomial approaches as shown in Table 3. The unbalanced binomial network has the highest accuracy and lowest SSE and RMSE value. This is misleading however as the classifier could classify everything as 0 to yield an accuracy value equal to the proportion of 0 or "no" in the actual values which in this case is 93.75%. A similar situation arises with the SSE and RMSE figures. The CV and R^2 terms are the only error metrics presented which can be used to effectively compare the models as they are dimensionless.

Table 3: Testing results – binomial, balanced versus unbalanced data

Classifier	**Testing Data**				
	Accuracy	SSE	RMSE	CV	R^2
Unbalanced binomial EN[1]	93.79	57.01	0.19	719.52	0.08
Unbalanced binomial EN[2]	88.75	100.79	0.28	343.34	0.15
Balanced binomial EN[2]	60.52	60.05	0.48	94.16	0.55
Unbalanced binomial NN[1]	93.750	38.850	0.193	837.141	0.053
Unbalanced binomial NN[2]	88.792	67.799	0.285	349.754	0.109
Balanced binomial NN[2]	63.264	38.235	0.476	95.169	0.544

[1]Results show the mean values from the 16 individual binomial models for household categories A-P

[2]Results show the mean values from the 8 binomial models for household categories B, C, F, G, H, I, K and M, i.e., the households used in the balanced data analysis. The six month time frame is used.

For comparison, we note that the corresponding balanced binomial NN using a single work week window produced Accuracy, SSE, RMSE, CV and R^2 values of (57.017, 41.798, 0.493, 98.676, 0.513). These can be compared with the last line of Table 3 which shows the values of the six month window. Such empirical evidence was used to guide the selection of the design parameters during model evaluation.

Details of the performance of the balanced data binomial models are shown in Table 4.

Table 4: Testing results - household category binomial models using balanced data.

Classifier	**Household category**	**Test Data**				
		Accuracy	SSE	RMSE	CV	R^2
Bal. Bin. EN	B	58.73	14.19	0.47	87.94	0.58
Bal. Bin. EN	C	44.70	66.56	0.50	90.14	0.55
Bal. Bin. EN	F	62.92	22.17	0.50	105.77	0.47
Bal. Bin. EN	G	70.35	42.03	0.46	101.62	0.53
Bal. Bin. EN	H	57.14	37.86	0.51	105.07	0.47
Bal. Bin. EN	I	55.94	180.27	0.49	99.05	0.51
Bal. Bin. EN	K	76.1	73.37	0.41	87.17	0.64
Bal. Bin. EN	M	61.76	38.89	0.48	82.13	0.61
Bal. Bin. EN	**Mean**	**60.96**	**59.42**	**0.48**	**94.87**	**0.55**
Bal. Bin. NN	B	50.00	10.46	0.51	102.29	0.48
Bal. Bin. NN	C	63.79	39.99	0.48	95.89	0.54
Bal. Bin. NN	F	70.69	13.45	0.48	96.30	0.54
Bal. Bin. NN	G	64.39	30.49	0.48	96.13	0.54
Bal. Bin. NN	H	54.17	25.41	0.52	102.90	0.47
Bal. Bin. NN	I	57.63	119.81	0.49	98.10	0.52
Bal. Bin. NN	K	79.37	41.98	0.38	76.62	0.71
Bal. Bin. NN	M	66.07	24.28	0.47	93.13	0.57
Bal. Bin. NN	**Mean**	**63.26**	**38.24**	**0.48**	**95.17**	**0.54**

For brevity we present just the results of the individual classifiers, i.e., asking whether test meters belong to a particular household category. The balanced binomial EN model has the highest R^2 value of 0.55 which signifies that 55% of the variability in the actual values is explained by the model. The best binomial NN was obtained using balanced data with an R^2 value of 0.54. Note the high performance for single adult household category K. This may indicate that category K is more distinctive than the other categories. Recall that category K accounts for 25% of the population, see Table 1. Table 5 shows a sample confusion matrix when testing sample meters for membership of household category K (single adult) using the best binomial NN. The matrix is diagonally dominant but more households are classified as true (163) than as false (123). In this example, the classifier is giving "false positives".

Table 5: Sample confusion matrix, household category K, binomial NN using balanced data

Test		**Predicted**		
		False	True	Σ
Actual	False	100	43	143
	True	23	120	143
	Σ	123	163	286

Scatter plots such as Figure 6 are useful to visualise the partitioning ability of the classifier. The *y*-axis refers to predicted probability (equivalent to the probability that meter belongs to a particular class). The *x*-axis labelled as "index" refers to the i^{th} test object. The data are evenly distributed data between the upper and lower halves of the plot area for both the EN and NN. Dark coloured dots represent households that are *true*, that is, test households that are in category K. Any dark dot above 0.5 is correctly classified. We see some dark dots below the 0.5 threshold. These are test objects that are incorrectly classified as not being in category K.

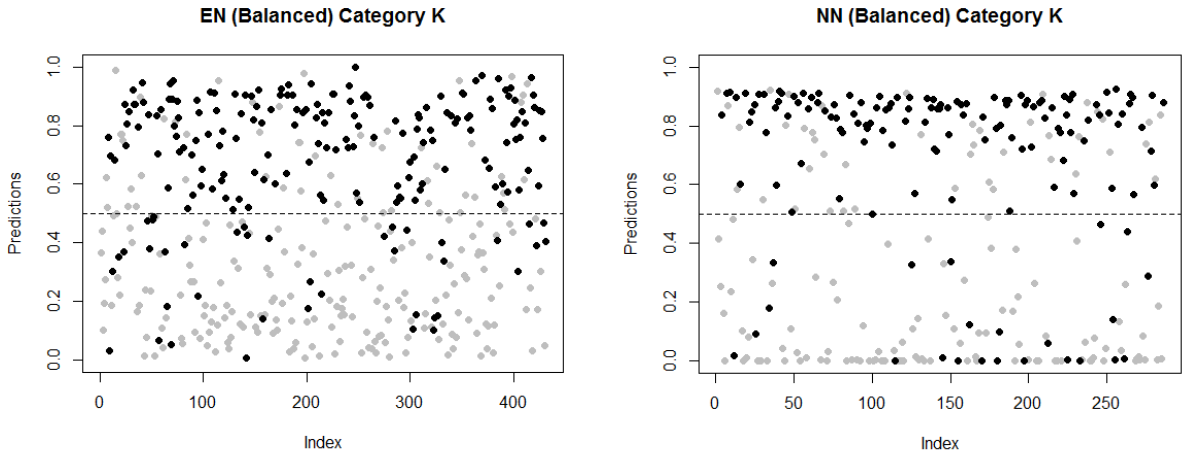


Figure 6: Scatter plot for household category K (single adult), binomial EN and NN using balanced data

The grey dots represent test objects that are *false*, that is, not in category K. Again we see that some are correctly classified (below the 0.5 line) and some are incorrectly classified (above the 0.5 line). These plots show that the classifiers have similar prediction accuracy for both *true* and *false*. The majority of the predictions are concentrated in the top and bottom quarters of the NN plot area as expected from a good classifier.

We had very limited success using multinomial classifiers using balanced data. Again the EN and NN had similar performances. The unbalanced data has the better R^2 value, but the balanced approach has the lower SCE and CV values as shown in Table 6 for the six month time frame. Recall that R^2 measures how well the variability in the actual values is captured by the model. If the model is distorted or biased towards a particular category then the R^2 value can also be distorted by the unbalance. The CV value is not affected by the imbalance as it only evaluates the relative closeness of the predictions to the actual values. It is useful when comparing models which use either balanced or unbalanced data.

Table 6: Testing results – multinomial approach

Classifier	**Testing Data**					
	Accuracy	SCE	SSE	RMSE	CV	R^2
Unbal. Multi EN	39.00	751.91	781.76	0.86	698.15	0.26
Bal. Multi EN	21.55	212.92	209.92	0.92	733.64	0.16
Unbal. Multi. NN	35.27	655.05	600.43	0.88	1409.23	0.22
Bal. Multi. NN	21.25	138.86	134.73	0.92	734.10	0.16

A sample confusion matrix for a multinomial NN is shown in Table 7. Category K can be predicted with 75% accuracy by the NN but category B displays an accuracy of 0%. The R^2 for this model was 0.16.

Table 7: Sample confusion matrix for multinomial NN using balanced data

Test		Predicted by Household category									
		B	C	F	G	H	I	K	M	Σ	% Acc
Actual by Household category	B	0	0	6	6	0	6	2	0	20	0
	C	0	0	4	10	1	3	1	1	20	0
	F	0	0	8	6	0	4	2	0	20	40
	G	0	0	5	2	1	8	4	0	20	10
	H	0	0	4	4	1	7	4	0	20	5
	I	1	0	1	2	0	8	8	0	20	40
	K	0	0	0	0	0	5	15	0	20	75
	M	0	0	10	4	0	3	3	0	20	0
		1	0	38	34	3	44	39	1	160	

5.2. Results Summary

In summary, the binomial approaches trained on the six month time series using balanced training data achieved the best performance. They are of less practical value than a multinomial classifier as they have to be tested against each household category and a weighted average calculated to yield an equivalent multinomial response. There was no significant difference between the EN and NN classifiers or simpler number of persons classification scheme. Some household categories were easier to identify than others. The R^2 value of the balance binomial NN for single person households was 0.71 (Table 4).

6. Discussion and Conclusion

This novel study describes an approach to household classification using smart meter data. The study presents a proof of concept for the use of ML and GLM models on new data sources such as SMD for use in the production of official statistics and by the public sector. The binomial approach

proved more useful to gain insight to household composition. The multinomial models have greater potential for practical applications but were less able to distinguish between the categories.

This study focused on exploring a specific consumer behaviour trial smart meter dataset with a view to learning a little more about it in the context of official statistics. The data were anonymised so could not be linked to other data sources but may yet provide a set of auxiliary information to allow NSIs determine whether a house is occupied or not, provide estimates of the number of persons living in a house. Rich data on households, giving a good indicator of a person or the number of persons living in a household, is highly valuable to developing small area statistics or census like statistics on small areas.

The CSO is exploring additional data sources, (Dunne, 2015). A building energy rating system has been in operation in Ireland since 2009. The CSO has access to this data but currently only one third of households have been rated. As this system evolves, it may be a potential administrative source that could be linked to live (un-anonymised) SM data to improve the classification performance. This project has not yet been costed, but is one of a number of possible data sources being considered for inclusion as a piece of the jigsaw in the developing National Data Infrastructure. While the CSO has access under the Statistics Act to access utility data such as SMD, it needs to evaluate how accessing such data can be socially justified and that any such access is proportionate and protects the privacy of the individual.

The insights gained during this study highlight some of the challenges and problems associated with classification schemes. The aim was to evaluate SMD to identify *existing* CSO household composition categories. As noted in Section 4.1, this smart meter study uses a simpler household categorisation system based on the numbers of adults and children sharing a dwelling unit. Some households, such as single person households (category K) appear to be more easily identified. However, the usage patterns for most existing CSO household categories are not sufficiently unique to be identified by their electricity consumption alone. There is a potential role for SMD driven models to estimate household composition in non-response or hard to reach households. It is also likely the classifiers could identify an empty (zero occupants) household. No such households were included in the CBT.

There is significant interest in NSI communities to identify and harness new data sources which may offer the opportunity for new insights. These sources may not be well structured, may have corrupt or missing segments, and may require considerable pre-processing to be manipulated into a useful format for analysis. Furthermore, the data may come from a domain not familiar to NSI staff and will involve a significant learning curve.

NNs and ML techniques have become more widely used for classification tasks, offering alternatives to traditional statistical for organisations intent on exploring new, possibly noisy, data sources. The NN and EN models had similar performance. There was no distinct advantage in favour of either the machine learning or generalised linear modelling approach. Neither approach was able to classify households with high reliability. The confusion matrices give some insight into how households can be mis-classified based on the similarity of usage patterns. Statistical models such as ENs may be more familiar to NSI communities in comparison to ML techniques so may be a more suitable approach.

Finally, in response to our research question whether CSO household composition can be estimated from analysis of SM electricity usage, we report only limited success in identifying households in general, but suggest that future studies linking SMD to supplementary information about the dwelling/building or other properties of the household could be beneficial.

Acknowledgement

We would like to thank the anonymous reviewers and Associate Editor for their constructive suggestions on early drafts of this manuscript.

Appendix 1: Features

The 21 features created for development of the models are described below. l is the total number of half hourly intervals over the particular time frame, n is the total number of intervals in a day, m is the total number of days in the time frame and E is the electrical demand in kWh:

1. Mean: mean consumption over the time frame l. $E_{mean} = \frac{1}{l}\sum_{i=1}^{l} E_i$
2. Max: maximum consumption during the time frame l. $E_{max} = \max(\{E_i\})$ where $1 \leq i \leq l$
3. ToUmax: the time slot i when *Max* occurs, $1 \leq i \leq n$. Note $n = l \pmod{48}$ as we cycle through the days in a time frame.
4. TEC: total electricity consumed over the time frame l. $E_{TEC} = \sum_{i=1}^{l} E_i$
5. MDM: Mean daily max is the average of the Max values for each of the m days. $E_{MDM} = \frac{1}{m}\sum_{j=1}^{m} E_j$ where $E_j = max\,(\{E_i\})$ for each m days: $1 + n(m-1) \leq i \leq nm$
6. Load Factor: This is the average of the ratios of the daily mean to daily maximum consumption. It is a measure of the peak of a household's load profile. A larger load factor indicates a household who uses electricity more evenly across the day while a low load factor indicates small periods of large consumption. For example, the load factor for the first day is $E_{LF_1} = \frac{(1/n)\sum_{i=1}^{n} E_i}{\max(\{E_i, 1 \leq i \leq n\})}$.
7. Variance: a measure of how far the electricity readings are spread out from the mean reading. $E_{VAR} = \frac{1}{l}\sum_{i=1}^{l}(E_i - E_{mean})^2$
8. Standard Deviation: a measure of the variation or dispersion around the mean reading, it is the square root of the variance. $E_{SD} = \sqrt{E_{VAR}}$
9. Range: the difference between the biggest and smallest electricity consumption readings. $E_{Range} = \max(\{E_i, 1 \leq i \leq l\}) - min\,(\{E_i, 1 \leq i \leq l\})$
10. Interquartile range (IQR): measures the difference between the third quartile and first quartile values of the data.
11. Morning max: the mean daily maximum electricity demand prior to 10am on a weekday. $E_{Mornmax} = \frac{1}{m}\sum_{j=1}^{m} E_j$ where $E_j = max\,(E_i)$ and i is within the first 20 time slots of each day.
12. Morning peak: the morning max minus the mean value between 10am and 12am on a weekday. This feature measures the size of a morning spike if one exists. It was observed that households with children were more likely to have a defined peak in the morning time on a weekday.
13. Morning range: the Morning max minus the minimum value before 10am.
14. Weekday Area: The area under the curve was approximated using the trapezoid rule.
15. Weekday Midpoint: The midpoint of the function was defined as half the total weekday area, the value returned was the time of day where the midpoint occurred.
16. Weekday Centroid: Analogous to the geometric centroid, the centroid of a function is the "centre of mass" of that function.
17. Weekday AM Slope: The slope of the early morning peak (up to 10 am) was taken as the rate of increase of energy consumption over time during the early morning period.

18 – 21. The procedures for features 15 – 17 were repeated to produce the equivalent features derived from the weekend energy consumption. These features were observed to differ to those during the working week, possibly due to behavioural changes at weekends.

Appendix 2: Error Metrics

The models were assessed using the following error metrics where y_i = predicted value of the i^{th} meter, t_i = true value of the i^{th} meter, N = Number of meters and $\bar{t}$ = mean of the true values.

1) Percentage of Correct Predictions (Accuracy): For the binomial model, the values predicted by the classifier are rounded to the nearest integer. For example, a prediction of 0.364 is rounded to 0, which indicates *false.* This says the meter does not belong to the category being tested. A predicted value of 0.759 is interpreted as *true* and means that the meter is assigned to that particular category. If the predicted category matches the true category, then the prediction is correct.

   For a multinomial classifier, a "winner-takes-all" approach assigns the category with the largest value to 1 and sets the remaining categories to 0. For example, a model concerned with four household categories produces output (0.25, 0.48, 0.10, 0.17). This is interpreted as (0, 1, 0, 0). This is a correct prediction if this was the true category of the meter.

2) Sum of Squared Error (SSE): The sum of the squared differences between the actual and predicted value. SSE = $\sum_{i=1}^{N}(y_i - t_i)^2$.

3) Root Mean Squared Error (RMSE): RMSE is an extension of SSE. The SSE is divided by the total number of meters to find the mean squared error (MSE). $RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - t_i)^2}{N}}$

4) Sum of Cross Entropy Error (SCE): The SCE is computed for each of the meters in the dataset and is summed over the entire dataset to get the SCE for the dataset. RMSE was used to compute the training error in the binomial classifier while SCE is used in the multinomial classifiers. SCE = $-1 \times \sum_{i=1}^{N}(t_i \times \log_{10} y_i)$.

   The reason for choosing SCE over RMSE for the multinomial classifier is demonstrated in the following example. A classifier concerned with four household categories produces (0.12, 0.57, 0.16, 0.15) and the true classification is (0, 1, 0, 0). The RMSE is 0.247 $\sqrt{(0.12-0)^2 + (0.57-1)^2 + (0.16-0)^2 + (0.15-0)^2} \div 1$.

   Now suppose the predicted output was (0.33, 0.57, 0.0, 0.1), then the RMSE is 0.304 $\sqrt{(0.33-0)^2 + (0.57-1)^2 + (0.0-0)^2 + (0.1-0)^2} \div 1$. Although the probability of the meter being classified as the second category is the same in both cases, the RMSE is different, 0.247 versus 0.304. Using SCE all but one of the error terms is zero and therefore the SCE for both cases is -1*1*log(.57) = 0.24.

5) Coefficient of Variation (*CV*): *CV* evaluates the relative closeness of the predictions to the actual values. The *CV* for a model describes the accuracy of the model in terms of the relative sizes of the residuals and the actual values. A high *CV* represents a large dispersion in the variables. An advantage to using this error term is that it is unitless and therefore it can be used to compare model performance. For balanced data, the *CV* value is just a multiple of the RMSE term but for unbalanced data it is particularly useful. $\frac{\sqrt{\frac{\sum_{i=1}^{N}(y_i - t_i)^2}{N}}}{\bar{t}} \times 100$

6) Pseudo R^2: R^2 quantifies how much of the variability is explained by the model. It indicates how well the data points fit some model representation of the data. Like *CV*, *R*-squared is unitless. In this study pseudo R^2 is defined as: $1 - \frac{\sum_{i=1}^{N}(y_i - t_i)^2}{\sum_{i=1}^{N}(t_i)^2}$

   The values of R^2 lie between 0-1. An R^2 value of 1 represents a perfect fit while a value of 0 represents inappropriate model fit.

References

Aydinalp, M., V. I. Ugursal, and A. S. Fung. 2002. "Modeling of the Appliance, Lighting, and Space-cooling Energy Consumptions in the Residential Sector Using Neural Networks." *Applied Energy* 71 (2): 87–110. DOI: http://dx.doi.org/10.1016/S0306-2619(01)00049-6.

Bousquet, O., and A. Elisseeff. 2002. "Stability and generalization." *Journal of Machine Learning Research* 2 (3): 499-526.

Breunig, M. M., H. P. Kriegel, R. T.Ng, and J. Sander. 2000. "LOF: identifying density-based local outliers." In *ACM sigmod record* 29(2): 93-104. DOI: http://doi.acm.org/10.1145/335191.335388.

CER. 2014. *Commission for Energy Regulation National Smart Metering Programme Smart Metering High Level Design*, Decision Paper CER/14/046. Available at: http://www.cer.ie/docs/000699/CER14046%20High%20Level%20Design.pdf (accessed March 2017).

Europea, Commissione. 2014. *A policy framework for climate and energy in the period from 2020 to 2030*. *COM (2014), 15*. Available at http://ec.europa.eu/smart-regulation/impact/ia_carried_out/docs/ia_2014/swd_2014_0015_en.pdf. (accessed March 2017).

CSO. 2011. "Census 2011 Reports", http://www.cso.ie/en/census/census2011reports/ (accessed September 2017).

Di Cosmo, V., S. Lyons, and A. Nolan. 2012. "Estimating the Impact of Time-of-use Pricing on Irish Electricity Demand." *The Energy Journal* 35 (2): 117-136. DOI: 10.5547/01956574.35.2.6.

Dunne, J. 2015. "The Irish Statistical System and the emerging Census opportunity." *Statistical Journal of the IAOS* 31(3): 391-400. DOI: 10.3233/SJI-150915.

Friedman, J, T. Hastie, and R. Tibshirani. 2009. "glmnet: Lasso and elastic-net regularized generalized linear models. R package version, 1(4)". CRAN: Wien, Austria. URL https://CRAN.R-project.org/package=glmnet.

Fritsch, S., F. Guenther, and M. Suling. 2012 "neuralnet: Training of neural networks. R package version 1.32. 2012." CRAN: Wien, Austria. URL https://CRAN. R-project. org/package= neuralnet.

Han, J. and M. Kamber. 2006. *Data Mining: Concepts and Techniques*. 2nd Edition, Morgan Kaufmann. ISBN 13: 978-1-55860-901-3.

Hand, D. J. 1998. "Data mining: Statistics and more?" *The American Statistician* 52 (2): 112-118. DOI:10.1080/00031305.1998.10480549.

He, H., and A.W. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 29 (9): 1263 – 1284. DOI: 10.1109/TKDE.2008.239.

Hickey, C., S. Kelly, P. Carroll, and J. O'Connor. 2015. "Prediction of Forestry Planned End Products Using Dirichlet Regression and Neural Networks". *Forest Science*, 61(2), 289-297. DOI: https://doi.org/10.5849/forsci.14-023.

Hopfield, J. J. 1984. "Neurons with graded response have collective computational properties like those of two-state neurons." *Proc. Natl. Acad. Sci.* 81(10): 3088-3092.

Hoerl, A. and R. Kennard. 1988. "Ridge regression." *Encyclopedia of Statistical Sciences* 8: 129-136. DOI: 10.1002/0471667196.ess2280.pub2.

ISSDA, Irish Social Science Data Archive; Commission for Energy Regulation (CER), Available at: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/. (accessed March 2017).

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, RC Team and M. Benesty, 2014. "caret: Classification and regression training. R package version 6.0–21". CRAN: Wien, Austria. URL https://CRAN.R-project.org/package=caret.

Liao, W. 2005. "Clustering of Time Series Data - a Survey." *Pattern Recognition* 38 (11): 1857–1874. DOI: http://dx.doi.org/10.1016/j.patcog.2005.01.025.

Marsland, S. 2009. *Machine Learning: An Algorithmic Perspective*. Chapman and Hall/CRC. ISBN:978-1-4200-6718-7.

McKenna, E., I. Richardson, and M. Thomson. 2012. "Smart meter data: Balancing consumer privacy concerns with legitimate applications." *Energy Policy* 41: 807-814. DOI: https://doi.org/10.1016/j.enpol.2011.11.049.

McLoughlin, F., A. Duffy, and M. Conlon. 2012. "Characterising Domestic Electricity Consumption Patterns by Dwelling and Occupant Socio-economic Variables: An Irish Case Study." *Energy and Buildings* 48: 240–48. DOI: http://dx.doi.org/10.1016/j.enbuild.2012.01.037.

Molina-Markham, A., P. Shenoy, K. Fu, E. Cecchet, and D. Irwin. 2010. "Private Memoirs of a Smart Meter." In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-efficiency in Building*, 61–66. DOI: http://dl.acm.org/citation.cfm?id=1878446.

Murphey, Y.L., H. Guo, and L.A. Feldkamp. 2004. "Neural Learning from Unbalanced Data." *Applied Intelligence* 21 (2): 117–128. DOI: http://dx.doi.org.ucd.idm.oclc.org/10.1023/B:APIN.0000033632.42843.17.

Newing, A., B. Anderson, A. Bahaj, and P. James. 2016. "The role of digital trace data in supporting the collection of population statistics–the case for smart metered electricity consumption data." *Population, Space and Place* 22 (8): 849–863. DOI: 10.1002/psp.1972.

Patidar, R., and L. Sharma. 2011. "Credit Card Fraud Detection Using Neural Network." *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, 2231–2307. DOI: http://dx.doi.org/10.1007/978-3-319-46675-0_53.

R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. (accessed March 2017).

Ripley, B. and W. Venables. 2011." nnet: Feed-forward neural networks and multinomial log-linear models. R package version, 7(5)". CRAN: Wien, Austria. URL https://CRAN.R-project.org/package=nnet

Seyb, A., R. McKenzie, and A. Skerrett. 2013. "Innovative Production Systems at Statistics New Zealand: Overcoming the Design and Build Bottleneck." *Journal of Official Statistics* 29 (1): 73-97. DOI: https://doi.org/10.2478/jos-2013-0005.

Statistics Act. 1993. *An Act to provide for the collection, compilation, extraction and dissemination of official statistics and for related maters*. Available at: http://www.irishstatutebook.ie/1993/en/act/pub/0021/print.html. (accessed March 2017).

Tibshirani, R. 1996. "Regression shrinkage and selection via the lasso." *J. R. Statist. Soc.B* 58, 267-288.

Van Gerwen, R., S. Jaarsma, and R. Wilhite, 2006. *Smart Metering, Technical Report* Available at: https://idc-online.com/technical_references/pdfs/electrical_engineering/Smart_Metering.pdf. (accessed 14/3/17).

Wang, X., K. Smith, and R. Hyndman. 2006. "Characteristic-based clustering for time series data." *Data Mining and Knowledge Discovery* 13: 335-364. DOI: http://dx.doi.org/10.1007/s10618-005-0039-x.

Zhang L., and B. Zhang. 1999. "A Geometrical Representation of McCulloch–Pitts Neural Model and Its Applications." *IEEE Transactions on Neural Networks* 10 (4): 925 – 929. DOI: http://dx.doi.org/10.1109/72.774263.

Zou, H., and T. Hastie. 2005. "Regularization and variable selection via the elastic net." *J. R. Statist. Soc. B* 68: 301-320. DOI: 10.1111/j.1467-9868.2005.00503.x.