Application of a clustering framework to UK domestic electricity data

Ian Dent, Uwe Aickelin, Tom Rodden

***Abstract*—The UK electricity industry will shortly have available a massively increased amount of data from domestic households and this paper is a step towards deriving useful information from non intrusive household level monitoring of electricity. The paper takes an approach to clustering domestic load profiles that has been successfully used in Portugal and applies it to UK data. It is found that the preferred technique in the Portuguese work (a process combining Self Organised Maps and Kmeans) is not appropriate for the UK data. The work uses data collected in Milton Keynes around 1990 and shows that clusters of households can be identified demonstrating the appropriateness of defining more stereotypical electricity usage patterns than the two load profiles currently published by the electricity industry.**

The work is part of a wider project to successfully apply demand side management techniques to gain benefits across the whole electricity network.

I. Introduction

The electricity market in the UK is subject to various pressures and is currently undergoing a period of major change. Some of these pressures are arising from UK specific situations, such as the history and current design of the National Grid, and others from worldwide trends, such as the need to reduce carbon emissions and the declining sources of hydro-carbon fuels. New technologies, such as electric cars and their need for household charging facilities, are expected to become much more prevalent. In addition, the drive to change the mix of electricity generation technologies to include more renewable technology, the desire to reduce carbon dioxide by switching non-electric demand such as gas central heating to the electricity network, and the impact of climate change, with its associated change in electricity demand for cooling or heating and more frequent extreme weather events, will impact on the market.

An important factor influencing the UK electricity market is that the presumption by consumers of the availability of an infinite supply of electricity, albeit at a cost, is no longer valid and domestic users will have to adapt to changing approaches to using electricity or suffer from increasing unreliability of the supply. [3] provides insight into the concerns of the industry in the USA and a number of these also apply to the UK market.

Corresponding author is Ian Dent (phone: +44 115 846 6568; email: ird@cs.nott.ac.uk), University of Nottingham, School of Computer Science, Jubilee Campus, Nottingham, NG8 1BB, UK

Professor Uwe Aickelin is with the Intelligent Modelling and Analysis Group, University of Nottingham, (email: uwe.aickelin@nottingham.ac.uk).

Professor Tom Rodden is with the Horizon Digital Economy Research Institute, University of Nottingham, (email: tar@cs.nott.ac.uk).

Prior to the planned roll out of smart meters, electricity suppliers were reliant on a single meter reading (or possibly 2 readings for households with Economy 7 meters) giving total usage for a 3 monthly period. There was no information on times of electricity usage, both time of day (beyond the Economy 7 period), and day by day. Electricity suppliers were therefore unable to offer tariffs to change user behaviour as there was no knowledge of the detailed behaviour.

The information available to monitor and to manipulate electricity usage will grow very rapidly, particularly with the roll out of Smart Meters which is planned to be complete in the UK by 2019.

[2] shows that the provision of Smart Meters will allow greatly increased analysis of a customer's electricity usage and provide the ability to make customised offers on pricing and supply availability. This will offer an opportunity to change customer behaviour (for example, to minimise usage during peak periods) or to increase efficiencies in the electricity supply chain in meeting the predicted demand [9].

The identification of typical electrical usage patterns for households is necessary as a starting point for:

- Defining the type of Demand Side Management program (e.g. peak clipping) to undertake to match the overall electricity supply goals.
- Assessing the impact of any initiatives to reduce overall energy usage in order to discover the amount of overall reduction which occurs during different times of the day.
- Allowing accurate aggregation to provide a pattern of total demand to be met by supply side generation and transmission.

Previous detailed monitoring research (for example [10]) has generally concentrated on working with a small number of households which are well understood, which include many different monitoring devices, and where the householder is supportive of the research and is prepared to dedicate time and effort to correct labelling of devices and to following researcher defined procedures. There remain a large number of households without a commitment to "green issues" and where detailed monitoring will not be possible, either due to lack of support from the householder, or for financial or time reasons.

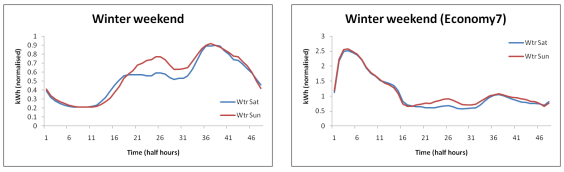
The paper describes work which forms part of a "demand side maximisation" project and focuses on identifying typical usage profiles for households and then clustering them into a few archetypical profiles with similar kinds of customers grouped together. Differences between an individual household profile and that of others within the same group can be used to suggest energy usage behaviour changes to reduce

overall electricity usage or to improve electrical efficiencies, possibly by time shifting the usage of particular appliances. In addition, particular groups (for example, large users during peak times) can be identified for targeting for reduction initiatives. The work tests the applicability of applying the framework defined by [6] to UK specific data and identifies possible enhancements or modifications to the framework in order to better fit the UK situation. In particular, the conclusion that a 2 stage process (Kohonen Self Organising Map and then Kmeans clustering) is the best approach to clustering the data is tested against the UK dataset.

This is the first step in exploring the limits of the information that is obtainable by non-intrusive monitoring at the whole household level. As well as the obvious overall usage information, future research will investigate knowledge that can be derived from the overall shape of the usage pattern as well as analysis of motifs that may repeat in the stream of usage data.

II. Background

The Electricity industry defines a process [5] for defining the details of eight standard usage profiles for the UK. The profiles take into account the season and the day of the week but only two are for domestic properties. As an example of the standard profiles, Figure 1 shows the profiles for the winter for Saturday and Sundays, both for Economy 7 customers and standard customers, plotted as 48 half hourly readings over the day. Economy 7 is a tariff offer that provides much cheaper night time electricity (typically for 7 hours between 11pm and 8am) at the expense of slightly increased day time charges.



(a) Standard users (b) Economy 7 users

Fig. 1: Example industry standard profiles

Figueiredo [6] takes various differing clustering approaches and reaches the conclusion that a combination of Self Organised Maps (using a 10 x 7 grid), followed by a Kmeans algorithm to reach a final set of 9 clusters, is the best approach as measured by a "cluster quality" measure defined in the paper.

The Kmeans algorithm requires a number of clusters as an input parameter (n) and works by randomly selecting an initial n locations for the centres of the clusters. Each data point is then assigned to one of the centre locations by selecting the centre that is nearest to that data point. The Kmeans method uses Euclidean distance calculated for centre $c = (c_1, c_2, ..., c_n)$ and point $p = (p_1, p_2, ..., p_n)$ as

$$distance = \sqrt{\sum_i (c_i - p_i)^2} \quad (1)$$

Once all the data points are assigned, each collection of points is considered, the new centre of the allocated points is calculated and the centre for that cluster is reassigned. The points are then reallocated to their new nearest centre and the algorithm continues as before until no changes are made to the allocations of points for an iteration. The method is highly dependent on the initial random allocation of centres [7].

The Self Organising Map (SOM) is a neural network algorithm that can be used to map a high dimension set of data into a lower dimension representation. In the work presented in this paper, the mapping is to a 2 dimensional set of representations which are arranged in a hexagonal map. Each sample (mean load profile for a given household) is assigned to a position in the map depending on the closeness of the sample to the values of the nodes assigned to each position in the map (using a Euclidean measure of distance). Initially the nodes are assigned at random but, as samples are assigned to the nodes, the node incorporates the assigned data. Over time, the map produces an arrangement where similar samples are placed closely together and dissimilar are placed far apart [8].

The Figueiredo approach includes the following stages:

- Cleaning of the data in order to cope with missing data and outliers in the data.
- Normalisation of the data to make differing readings comparable.
- Splitting of the data into typical types of day such as weekday, weekend, or season.
- Creation of representative daily load profiles. Figueiredo uses the mean across all available days within the type of day and season.
- Application of a number of clustering techniques in order to group the data into a pre-defined number of clusters and then the definition of a representative load profile for each cluster. A target number of clusters of nine is selected based on advice from the Portuguese electricity industry together with some investigation on the quality of the clusters obtained when trying numbers of clusters between 6 and 14.
- Calculation of the Mean Index Adequacy (MIA) as defined in [1] in order to assess the comparative suitability of the generated clusters.

Figueiredo makes use of Portuguese data on 165 consumers, with readings taken at a 15 minute frequency, in order to validate the approach taken.

The data used in this study is from an area of Milton Keynes, UK and was originally collected in 1988-91 by [4] but was stored on floppy disks which deteriorated physically and some of the original data was lost. The original data disks were rescued and, where possible, regenerated by Steve Pretlove of UCL and, more recently, by Alex Summerfield with the work detailed in [11]. The datasets have been made available in the UKERC data store.

III. Methodology

The approach detailed by Figueiredo [6] has been applied to the UK data as closely as possible in order to assess the suitability of the framework to the UK data. The individual steps in the process are detailed below.

A. Cleaning

Some of the UK data readings are missing readings for some hours of the day, either due to the way in which the data was recovered from floppy disks, or because of issues with the original collection of the data. For an initial view of the data, all the days which contained a missing hourly reading were omitted. Alternative approaches to replacing some of the missing data making use of available data from a similar day will be investigated in the future.

B. Normalisation

The UK data has been normalised within each day's readings by scaling all readings using the maximum hourly reading on the day set to 1. Thus all hourly readings are in the range 0-1. The effect of this normalisation is to focus on the shape of the usage pattern and not on the total usage. Two households with a similar shape but with differing total usages (e.g. if one household is much larger than the other) will have the same normalised load profile once scaling is done. The households will be clustered as similar in the further analysis whereas, depending on the way "similar" is defined, it might not be the intention to group these together (for example, if total electricity usage is to be the main differentiation between households).

C. Stratifying the data

The UK data was stratified using a split between weekend (Saturday and Sunday) and weekdays. It was further stratified into winter (the months of November, December, January, February, March, and April) and summer (the remaining months). With the variability of the UK climate, it may be more accurate to stratify the data based on daily temperatures rather than on the season and this will form the basis for future work. The data for winter weekends was arbitrarily chosen for further exploration as detailed in the remainder of this paper. Future work will concentrate on the other stratifications (e.g. summer weekday) which can be analysed in the same way. How individual households are allocated to the same or different groupings as the season or type of day changes will be investigated.

The Milton Keynes data has varying amounts of valid data for each household depending on the success of regeneration of the data after its rescue from floppy disks. The winter weekend data consists of between 25 and 111 valid days of readings for each of the households with a mean of 95 valid readings per household. Future analysis may suggest excluding some of the households with low values for valid data from the clustering but all have been included in this initial investigation.

D. Creation of load profiles

Each household has a representative average load profile generated by calculating the mean value for each hourly reading across all valid readings for the winter weekend. Other methods of calculating a representative profile could be adopted but this analysis has duplicated the approach taken with the Portuguese work.

E. Application of clustering algorithms

The Figueiredo approach compares the Kmeans algorithm with both a self-organised map (SOM) using a 3 x 3 grid and also with a 2 stage process of first creating a SOM with 10 x 7 grid (i.e. 70 load diagrams) which are then clustered using the Kmeans algorithm. This approach has been replicated with the UK data although the volume of households is less (165 in Portugal, 93 in the UK) and hence the reduction in dimensions from the first SOM stage is not as great as with the Portuguese data.

The Kmeans clustering method relies on a random starting situation and requires the desired number of clusters to be input. In order to minimise the effects of the random starting point, the clustering algorithm was run 1000 times with differing random seeds. Examination of the results suggests that the large number of runs allows the same optimum solution to be found regardless of the starting random seed.

The within cluster sum of squares was calculated for each of the input numbers of clusters from 2 to 15. As the number of clusters increases, the total sum of squares will decrease (with the extreme example of each sample being in its own cluster with a total within cluster sum of squares being 0) and the graph (Figure 2) can be examined to find an obvious "elbow" that denotes an appropriate number of clusters to select. The graph can be seen to be fairly uniform with no obvious elbows apart from that at 3 and possibly that at 9. In order to match the Portuguese work, the input value of 9 clusters was used for future analysis.

F. Calculation of adequacy measure

A measure is needed for assessing the quality of the clusters generated in order that the differing approaches can be compared. A good clustering scheme will create clusters where the members of a particular cluster are closely grouped but where the differences between members of differing clusters (or the representative profiles for the clusters) are well separated. A measure, Mean Index Adequacy (MIA), is defined in [1] as

$$MIA = \sqrt{\frac{1}{K}\sum_{k=1}^{K} d^2(r^{(k)}, C^{(k)})} \tag{2}$$

where K clusters (k = 1..K) have been defined, $r^{(k)}$ is a load profile assigned to cluster k and $C^{(k)}$ is the calculated centre of the cluster k.

The distance between 2 load diagrams is defined as

$$d(l_i, l_j) = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(l_i(h) - l_j(h))^2} \tag{3}$$

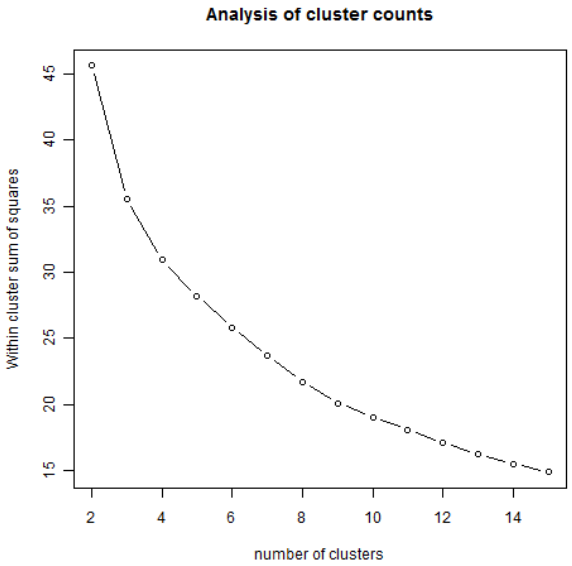


Fig. 2: Varying numbers of clusters input to Kmeans

where H is the number of individual readings in each load diagram (24 hourly readings) and $l_i(h)$ and $l_j(h)$ are the hth readings for two profiles, l_i and l_j.

The MIA can be better described as

$$MIA = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\sum_{r} d^2(r^{(k)}, C^{(k)})} \quad (4)$$

to signify the need to sum over all the distance calculations for each of the load profiles assigned to the given cluster (the distances between the load profiles and the cluster centre).

A lower value of MIA for a particular clustering solution signifies that the load profiles assigned to the calculated clusters are grouped closely together and hence a low value for MIA is better and shows more compact clusters. The measure is useful as a comparison between differing clustering algorithms (where a lower value shows more compact clusters) but has little meaning as an absolute value.

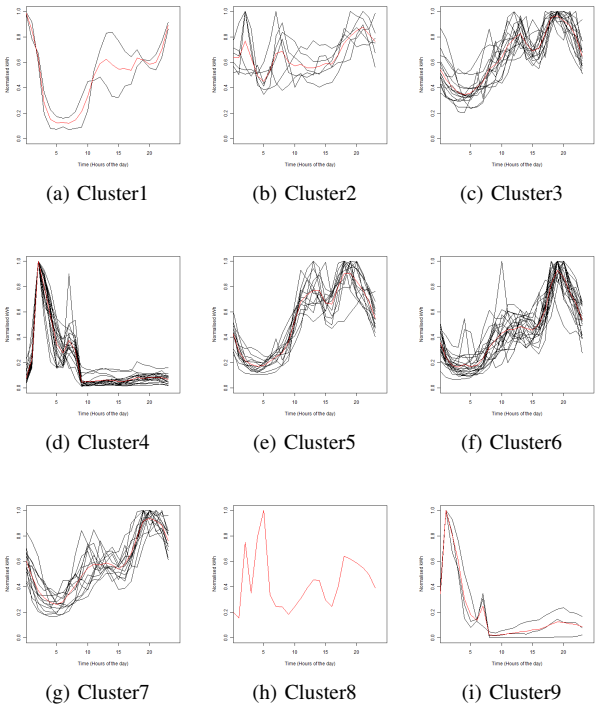
The analysis work used R 2.12.2 running on a Samsung R580 laptop with Windows 7 Enterprise 64 bit operating system Service Pack 1. The laptop used an Intel i3 CPU (M350) running at 2.27 GHz and contained 3GB of memory.

IV. Results

Differing clustering approaches were considered in order to explore the most appropriate for the UK data.

A. Kmeans

Initially the Kmeans clustering algorithm, with a target of 9 clusters, was used to form the clusters. The clustering results using the Kmeans algorithm can be seen in Figure 3 where the black lines show the load profiles for the individual households allocated to the particular cluster and the red line shows the calculated representative profile for the cluster (the centroid). Where only one household is allocated to a cluster (e.g. as with "Cluster8"), the red line is overlaid on the black line.



(a) Cluster1 (b) Cluster2 (c) Cluster3

(d) Cluster4 (e) Cluster5 (f) Cluster6

(g) Cluster7 (h) Cluster8 (i) Cluster9

Fig. 3: Clusters generated using Kmeans

The number of households allocated to each cluster by each technique are detailed in Table I.

TABLE I: Size of clusters

	1	2	3	4	5	6	7	8	9
Kmeans:	2	6	15	19	13	21	13	1	3
SOM:	2	6	15	22	12	9	12	1	14
2 Stage:	6	6	13	19	15	22	8	1	3

B. Self Organising Map

The Kohonen Self Organising Map algorithm was applied to the data using a hexagonal grid of 3 x 3 (i.e. 9 clusters). This creates the map of load profiles as shown in Figure 4.

Plotting the household load profiles alongside the calculated cluster centres produces the results in Figure 5 with the numbers of households allocated to each cluster listed in Table I. The clusters are numbered randomly and the order in the figure has been modified in order to match the Kmeans clusters as far as possible. The match between the generated clusters is visually obvious with the exception of "Cluster9".

C. Two stage process

The conclusion in [6] is that the application of a Kohonen Self Organising Map algorithm to the data in order to create 70 (10 x 7) clusters in a hexagonal grid followed by the application of the Kmeans algorithm to the SOM output produces

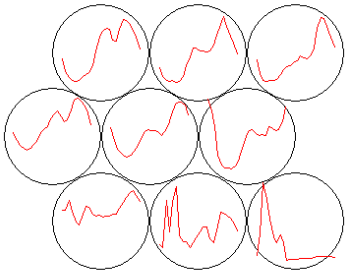
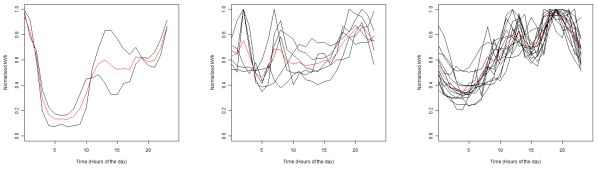
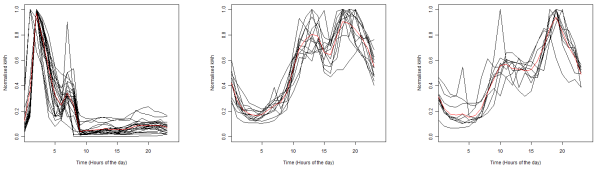


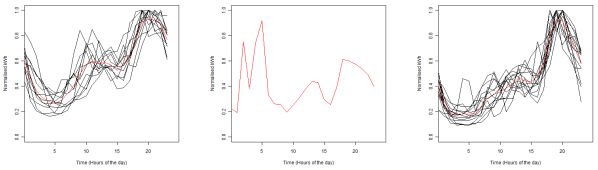
Fig. 4: Kohonen self organised map using 3 x 3 grid



(a) Cluster1 (b) Cluster2 (c) Cluster3



(d) Cluster4 (e) Cluster5 (f) Cluster6



(g) Cluster7 (h) Cluster8 (i) Cluster9

Fig. 5: Clusters generated using Kohonen Self Organising Maps

the best clusters as defined by the MIA measure. This work was replicated using the UK data although the number of households is lower than that used in the Portuguese work.

The intermediate map generated by the SOM is shown at Figure 6. The intermediate load profiles shown are then input to the Kmeans algorithm in order to generate 9 final clusters. The original allocation of household load profiles to the intermediate SOM and thence to the final clusters is then examined in order to determine the number of households in each final cluster and to allow for plotting of the final cluster profiles alongside the households allocated to that cluster. Again the order of the generated clusters has been altered to match the Kmeans generated clusters as closely as possible. The results are shown in Figure 7 with the number of households allocated to each cluster detailed in Table I.

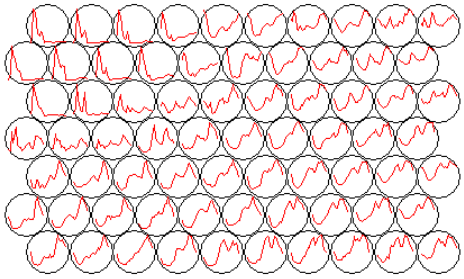
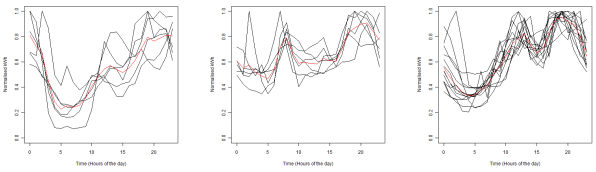
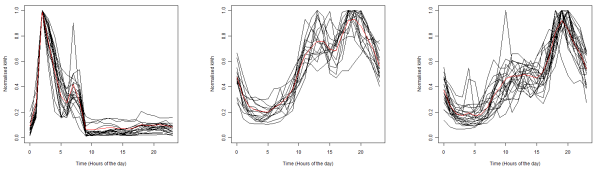


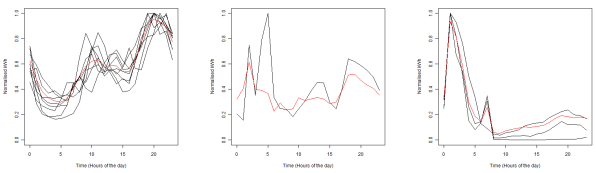
Fig. 6: Kohonen self organised map using 10 x 7 grid



(a) Cluster1 (b) Cluster2 (c) Cluster3



(d) Cluster4 (e) Cluster5 (f) Cluster6



(g) Cluster7 (h) Cluster8 (i) Cluster9

Fig. 7: Clusters generated using the 2 stage process

D. Comparison of clustering techniques

The MIA figures for each clustering approach are listed in Table II with a lower figure denoting more compact clusters. The results show the best algorithm for clustering (as measured by MIA) is Kmeans.

The MIA measure is very sensitive to the few profiles which differ from the profile for the generated cluster to which they are allocated. This sensitivity may detract from the MIA as a good measure of clustering success as, whilst most of the households may be well clustered, a single household profile allocated to one cluster rather than another can greatly increase the MIA value and hence reduce the measured effectiveness of the clustering. It is proposed in future work to examine alternative clustering measures and to assess the sensitivity of the measures to a few profiles which are difficult to allocate to clusters.

TABLE II: MIA calculations

	Kmeans	Kohonen SOM	2 stage process
MIA value:	0.3050533	0.3166297	0.3205487

The graphs showing the generated clusters and the households that are allocated to each cluster show that each technique produces some clusters which appear visually to be very similar but also some clusters that vary widely. In particular the "Cluster9" is significantly different for the various clustering techniques. The numbers of households allocated to each cluster vary and this demonstrates that the clustering techniques will have differing levels of success in generating the best split into clusters.

V. Conclusions and future work

The work demonstrates that UK domestic load profiles can be usefully clustered and the visual impression from the cluster representative profiles is of very differing shapes of usage. In particular, the load shapes differ significantly from the standard domestic profiles used by the industry which are only differentiated by Economy 7 usage (see Figure 1). This shows that the application of appropriate clustering techniques will allow for more accurate differentiation between household usage patterns than that currently published by the industry and will lead to more accurate representative profiles which can be used for demand aggregation, supply side planning, marketing and other purposes.

The selection of nine as the target number of clusters reflects the decision taken in Portugal. The evidence for selecting nine clusters for the UK winter weekend data is weak and more investigation of an appropriate target number of clusters appropriate to the UK data is planned.

The work undertaken in Portugal using Portuguese data concluded that using a two-stage process of building a Self Organising Map and then applying a Kmeans clustering algorithm was the most effective in generating well distinguished clusters as measured by the MIA measure. The UK data does not support this conclusion and the best MIA figure is from the simple application of the Kmeans algorithm. In fact, it was found that the SOM technique alone provided better results (as measured by the MIA measure) than the two-stage process.

Analysis has been concentrated on the winter weekend data and other slices across the data may show differing results. In particular it may be found that households are clustered together differently for different types of day (by season or weekend/weekday) and year long stable clusters, with the same members for each season, may not be identifiable. Future work is planned to investigate this further.

The MIA measure of the quality of the generated clusters is very sensitive to a few households which are hard to allocate and differing measures of cluster quality will be investigated in the future.

The normalisation used in the exercise has the effect of comparing shapes of usage but not absolute values of usage and a clustering approach that differentiates a household using much more electricity from another using less may be required (depending on the use to be made of the clusters found). The appropriateness of the normalisation is related to the definition of "similar" users which will be explored in future work.

Acknowledgements

This data was accessed through the UK Energy Research Centre Energy Data Centre (UKERC-EDC). Our acknowledgements to the Building Research Establishment, which provided access to the original 1990 data set from Milton Keynes Energy Park, and to Bartlett School of Graduate Studies, University College London for processing and cleaning the raw data.

This work is possible thanks to EPSRC grant reference EP/I000496/1.

References

[1] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader. Customer characterization options for improving the tariff offer. *Power Systems, IEEE Transactions on*, 18(1):381–387, 2003.

[2] DECC. Towards a smarter future, government response to the consultation on electricity and gas smart metering. 2009.

[3] US DOE. Grid 2030: A national vision for electricity's second 100 years, 2003.

[4] J. Edwards. Low energy dwellings in the Milton Keynes Energy Park. *Energy Management*, 26:32–33, 1990.

[5] Electricity Association. Load profiles and their use in electricity settlement. *UKERC*, 1997.

[6] V. Figueiredo, F. Rodrigues, Z. Vale, and J.B. Gouveia. An electric energy consumer characterization framework based on data mining techniques. *Power Systems, IEEE Transactions on*, 20(2):596–602, 2005.

[7] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. 1988.

[8] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 2002.

[9] Ofgem. Project discovery - options for delivering secure and sustainable energy supplies, 2010.

[10] I. Richardson, M. Thomson, D. Infield, and C. Clifford. Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings*, 42(10):1878–1887, 2010.

[11] AJ Summerfield, RJ Lowe, HR Bruhns, JA Caeiro, JP Steadman, and T. Oreszczyn. Milton Keynes Energy Park revisited: Changes in internal temperatures and energy usage. *Energy and Buildings*, 39(7):783–791, 2007.