Pattern recognition and segmentation of smart meter data

Barry McDonald[1] Peter Pudney[2] Jia Rong[3]

(Received 11 February 2013; revised 22 May 2014)

Abstract

In Australia, Smart Meters automatically provide electricity suppliers with half-hour energy use data for each customer. This data can be used to classify customers into different categories. To this end, electricity supplier AGL provided MISG participants with data from 772 anonymous Victorian customers, collected between 2011-07-16 and 2012-01-30, and the corresponding series of half-hour temperature readings for Melbourne. The goals were to identify a small number of load profiles that could be used to classify customers, and to identify which customers have significant cooling loads and which customers have significant heating loads. For each customer there was a time series of 9552 half-hour periods, which made the dimensionality of the problem too high for cluster analysis of the entire sample data. Therefore analysis proceeded in two phases. First, the data were explored using various

http://journal.austms.org.au/ojs/index.php/ANZIAMJ/article/view/6743 gives this article, © Austral. Mathematical Soc. 2014. Published May 27, 2014, as part of the Proceedings of the 2012 Mathematics and Statistics in Industry Study Group. ISSN 1446-8735. (Print two pages per sheet of paper.) Copies of this article must not be made otherwise available on the internet; instead link directly to this URL for this article.

methods of data visualisation, including time series plots, scatterplots and heatmaps of electricity use against temperature and time, Fourier series analysis and load duration curves. Exploration suggested that some automatic data-selection rules would be useful, for example to eliminate premises with long periods of zero electricity, presumably due to vacancy. Based on the data exploration, summary statistics were chosen that would represent each customer, and these were used in the next phase, cluster analysis. Second, three approaches were used for clustering: self-organising maps, agglomerative clustering, and K-means clustering. Each of these methods produced interpretable clusters indicating different types of customer. Agglomerative clustering with complete linkage was good for picking out small very distinctive clusters, and Ward's linkage also performed well provided sufficient clusters were allowed. Computational limitations mean these two techniques cannot be directly used on very large samples—AGL has hundreds of thousands of customers. However, the cluster centroids from a pilot study, such as the sample provided to MISG, could be used as initial estimates for feeding into K-means clustering, providing the twin benefits of interpretable clusters and computational efficiency.

Subject class: 62P30
Keywords: Pattern recognition, cluster analysis, segmentation, smart meters

1 Introduction

Traditionally, the electricity used on a premises is measured by a meter that accumulates total electrical energy use since the meter was installed. To read these simple electromechanical or electronic meters, electricity suppliers must visit the meter, typically quarterly. Meanwhile, electricity retailers settle their accounts with the market operator weekly, without knowing until months later exactly how much electricity their customers have used.

Simple meters are starting to be replaced by 'Smart Meters' that introduce new functionality, including the ability for electricity suppliers to read meters remotely, interfacing with in-home displays that provide electricity users with information about electricity use and pricing, and measuring energy use in half-hour intervals. Eventually, Smart Meters will also allow the retail price of electricity to vary with time of use (in half-hour intervals). This could in turn allow the retail price to better reflect the instantaneous cost of generating and transmitting electricity, which varies significantly with the total demand for electricity. The possible uses of Smart Meters have received growing attention in recent academic literature [1, 2, 3, 4, 5, 6, 7], as well as lay literature.

AGL has started collecting half-hour energy use data from customers with Smart Meters in Victoria and New South Wales. The data available for MISG included data collected between 2011-07-16 and 2012-01-30 from 772 Victorian customers, giving a time series of 9552 half-hour readings for each customer. Each meter reports the energy imported by the premises during each half-hour period and, if the premises has photovoltaic panels for collecting solar energy, the energy exported by the premises during each half-hour period. Subtracting the export energy from the import energy gives the net energy use, which may be negative during periods when export energy exceeds import energy.

For most customers, energy use is dependent on heating or cooling loads. AGL also supplied half-hour temperature data for Melbourne.

AGL would like to use the Smart Meter data to:

- improve their forecasts of demand, so that they can manage their wholesale energy portfolio better;
- identify their most profitable customers;
- classify customers for marketing purposes.

The questions for MISG were:

- could we identify a small number of load profiles that could be used to classify customers?

- could we identify which customers have significant cooling loads and which customers have significant heating loads?

Section 2 details some exploratory analysis that was undertaken to visualise the data and identify relevant patterns and factors. Aspects of time and temperature are considered, along with initial modelling using spectral analysis and Load Duration curves. Section 3 introduces the ideas of cluster analysis, taking the example of self-organising maps, and demonstrates how even a limited look at the data can provide meaningful clusters. Section 4 details an approach based on stratifying the data into 148 combinations of time, temperature range, and weekday/weekend, and summarising each customer by their mean power use across these combinations. The pros and cons of various clustering methods are considered, including Complete clustering, Ward clustering, K-means clustering and others. Some overall conclusions are given in Section 5.

2 Data exploration and visualisation

2.1 Temporal patterns

The first step was to organise and visualise the data. It was useful to graph the data from some individual customers, to get a sense of the types and range of variation in electricity use. Individual data can show unexpected patterns—large for some customers and small for others—that could be missed by looking only at data that has been averaged over individuals.

Figures 1(a) and (b) show the first 21 days of power consumption for two apparently typical household customers, one high-use and one low-use. Note there are 336 half-hour periods in a week. The customer in Figure 1(b) perhaps uses gas for heating and cooking, and electricity only for lighting and appliances, as the electricity use is about a third that of Figure 1(a). In both graphs power use never drops to zero, probably because of appliances such

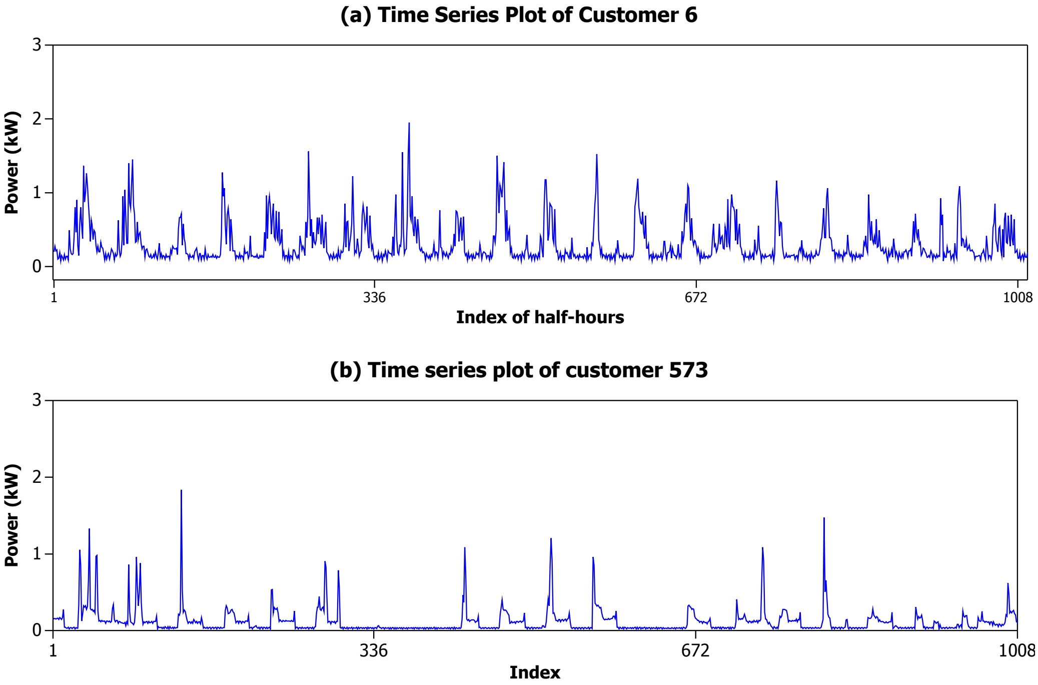

Figure 1: Typical (a) high-use and (b) low-use domestic customers.

as refrigerators or freezers. Figure 1(a) shows more consistent periodicity, with (usually) high peak evening use falling to low power use at night, but the pattern is slightly complicated by a weekend/weekday effect. Figure 1(a) starts with the first time period (00:00–00:30) on a Saturday, and there is a fair amount of electricity being used all through the daylight and evenings of the first Saturday and Sunday. However, on the next three days (Monday–Wednesday) there is a brief spike in consumption at breakfast time followed by a drop in consumption until evening, presumably because the occupants are away at work. This pattern (breakfast peak, low, big evening peak, nightly low) is a common pattern of consumption. The following Thursday–Saturday power use was high throughout the day, but from the Sunday–Friday of the following week the pattern of morning and evening peaks prevailed. In

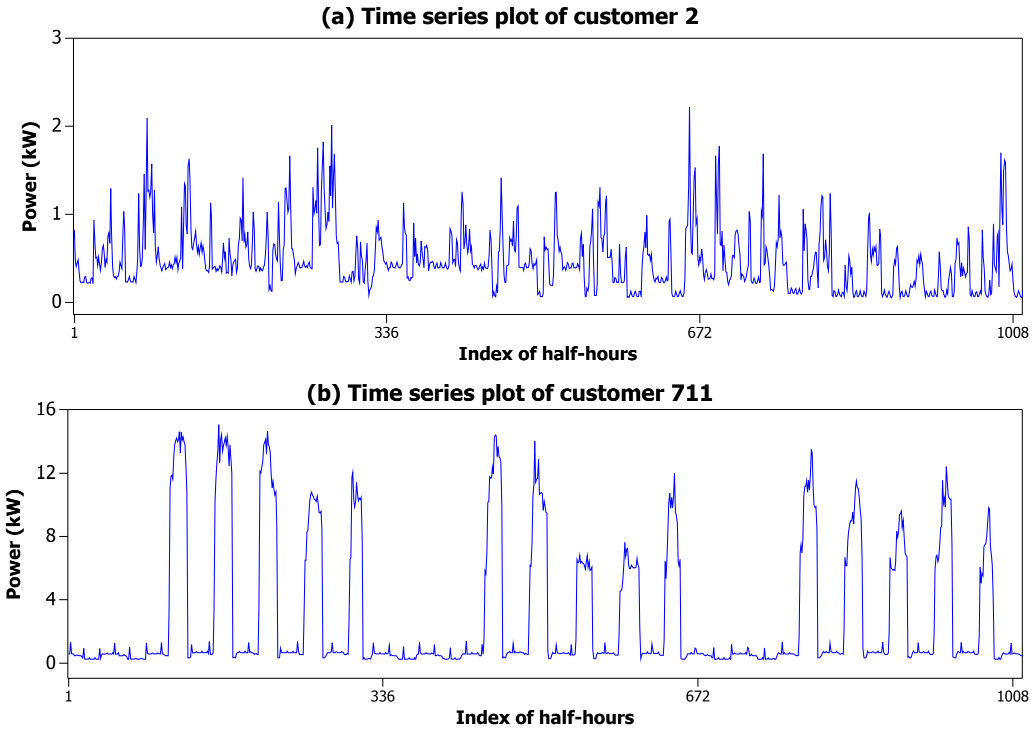

Figure 2: (a) Possibly a customer using electric heating, (b) High consumption during week days (probably a business).

Figure 1(b) the patterns are not as clear, although there is some evidence of morning and evening peaks in the middle of the graph. Figure 2(a) is a customer with similar electricity consumption and pattern as Figure 1(a), except that there are extended periods when the minimum energy use is above 0.25 kW; this is presumably because of heating, since these are winter days. We wish to distinguish patterns like Figures 1(a),(b) from 2(a).

Figures 2(b) and 3(a) and (b) show more unusual customers, who should perhaps be screened out first before applying a statistical clustering technique. Unfortunately only Figure 2(b) was regularly identified during the MISG.

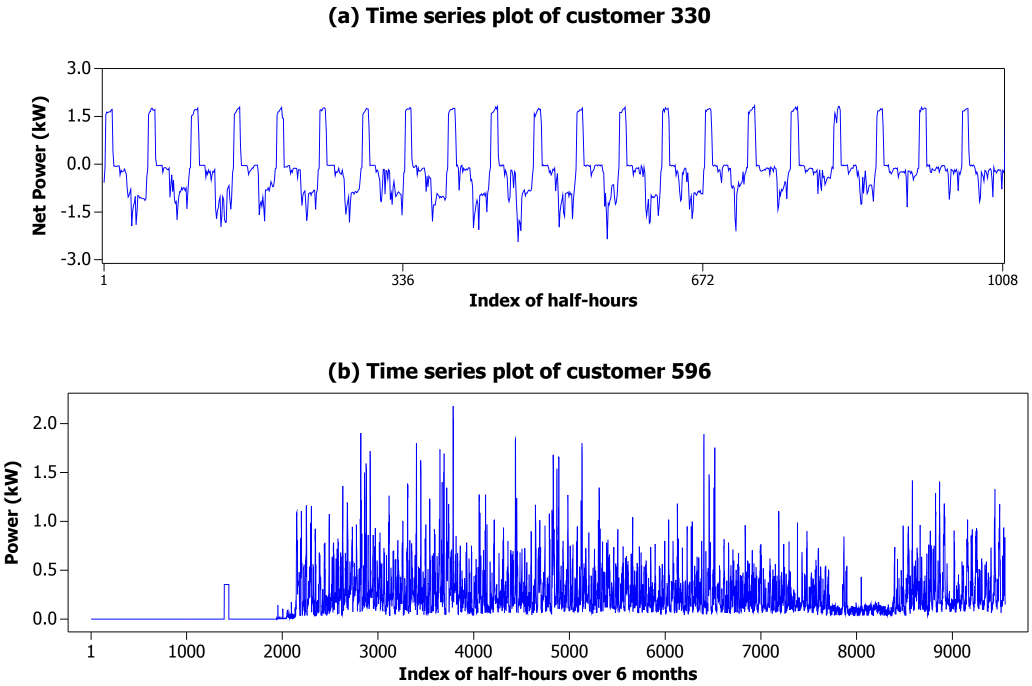

Figure 3: (a) Net energy consumption by a customer with solar panels, (b) Example of household with large variation over the 199 day period.

Figure 2(b) shows a consistent pattern of very high electricity use during weekday business hours, and very little electricity use on weekends. This commercial customer should not have been included in the study data, which focussed on residential customers. However, it is an example of a real outlier that could be used to develop automatic exclusion rules. There were residential customers with a more or less regular business-hours pattern of usage but much lower consumption. Some may be home businesses as distinct from commercial premises—the dividing line between types of customer is blurred. Figure 3(a) shows the first 21 days of net power usage for a customer with (presumably) photovoltaic cells (PV) for solar power generation. This customer had consistently high electricity consumption in the morning, perhaps due to heating on a timer, then during daylight their net electricity consumption became negative but highly variable as they exported energy to the grid. The positive and negative flows were represented by two separate columns in the database for these customers, and it was not initially clear during MISG what the interpretation was for the second column and what to do with them. So some MISG participants studied only the positive columns (power inflows). In retrospect it might have been better to have treated the PV customers as a separate population to be studied based on their net power consumption. However, there were only a few of them, and since MISG was essentially a scoping exercise it was decided to leave them in. AGL knows who their PV customers are without reference to their consumption records.

Figure 3(b) shows a customer whose behaviour varied considerably over the six months of data. The electricity seems to have been totally switched off for the first six weeks (apart from one brief episode), and then the customer appears to have been absent (probably on holiday) for a period around the 8000 mark on the horizontal axis (specifically, absent 24 December–6 January apart from brief returns). It is unclear how to handle these absences in the data, and it really comes down to choice by AGL. AGL may wish to detect and flag customers with extended periods of zeros, as being unusual customers, or perhaps it indicates a change in ownership or tenancy at the dwelling. Perhaps these customers should be removed from the data before using a statistical clustering technique. It is inconvenient to treat the zeros as missing data, since most multivariate clustering techniques require complete data. As for holidays, since many customers go away on holiday at different times of the year, these periods of low power use should in principle be considered part of the regular pattern of variation in customer usage. However, the periods of absence do add variance to the customer, and so make it harder for statistical routines to cluster the customers into interpretable groups.

These considerations suggest that some simple automatic rules should be devised for pre-selecting cases before applying statistical clustering techniques to the data. This pre-selection step is justified on the grounds that clustering techniques are tuned to detect subtle effects rather than gross differences:

if an effect is huge and has an obvious explanation, then we do not need a p-value or a dendrogram to tell us so.

2.2 Variation in electricity use by weeks

Figure 4 displays the weekly variation in electricity usage for some individual customers. The horizontal axis represents time of the week, in half-hour periods, and is shaded to indicate day (blue) and night (grey). The top half of the graph shows the mean electricity use in each half hour period of the week, averaged across the six months of the data. The bottom half of the graph is a heat map showing the energy use in (time, temperature) bins, where the vertical axis represents temperature. High energy use is represented by bright red, low energy use by pale pink. The histograms on the right of each diagram show total energy in each temperature bin (dark grey) compared to the relative frequency of the temperature (blue).

The top graph clearly corresponds to a business customer with high usage each day and virtually no usage at night. During the day the heatmap is almost uniformly bright red from bottom to top, indicating usage is independent of temperature. The other three customers are clearly residential. The second customer, for example uses very little electricity, and when they do use more (bright red) it is mostly in cooler weather near the bottom of the heatmap graph. The third customer has moderate use of electricity throughout the day and evening, and only a little more when cold or hot. The fourth customer makes heavy use of electricity for cooling during high temperatures (a lot of bright red at the top of the graph) and for heating when cold. The 'ribbons' of bright red suggest that if appliances are turned on then they are kept on for quite a while. The third and fourth customer show some evidence of lower electricity use during the middle of the day, but it looks like there is usually someone home using electricity.

In summary, these graphs give a quick overview of each customer, and help visually identify interesting customers.

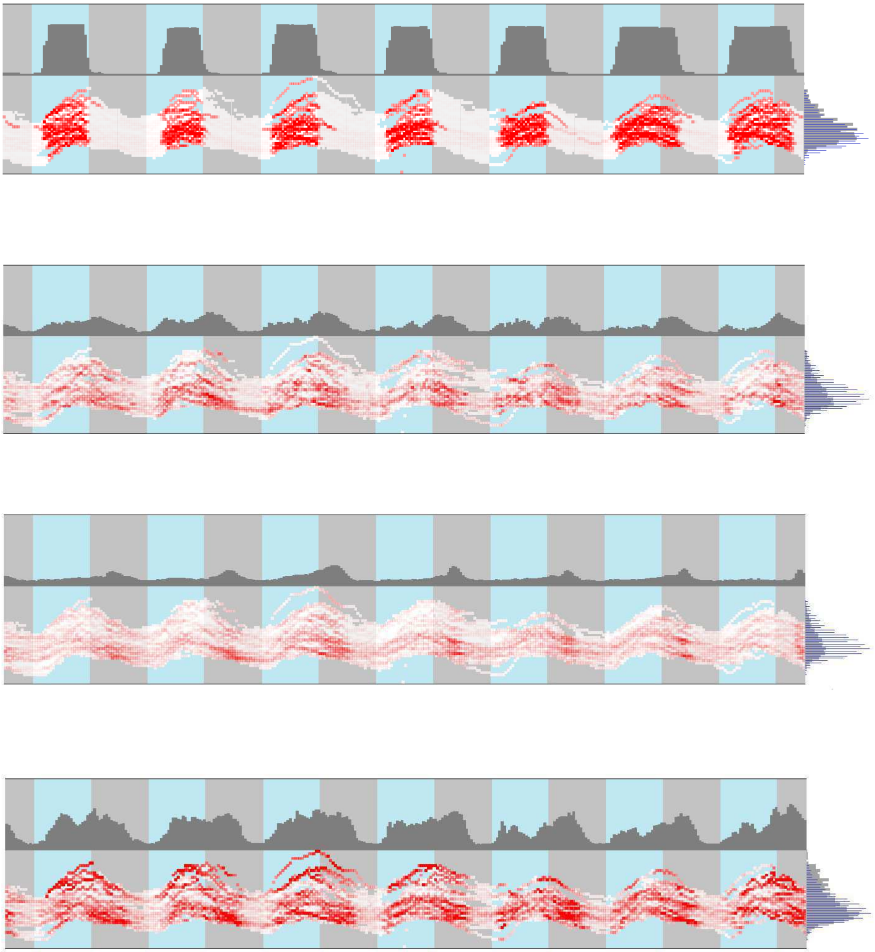

Figure 4: Visualisation of demand showing energy use by four customers, for half-hour periods throughout the week Monday-Sunday (horizontal axis). Histogram at top of each graph shows energy usage (0–1.2 kW) during 6 am–6 pm (blue background) and at night (grey background), while the heat map below additionally shows the variation with temperature (vertical scale from low to high temperature), with light colours meaning little energy use and redder meaning more energy use. The histograms at right show the relative frequency of temperatures (blue) and electricity use (grey).

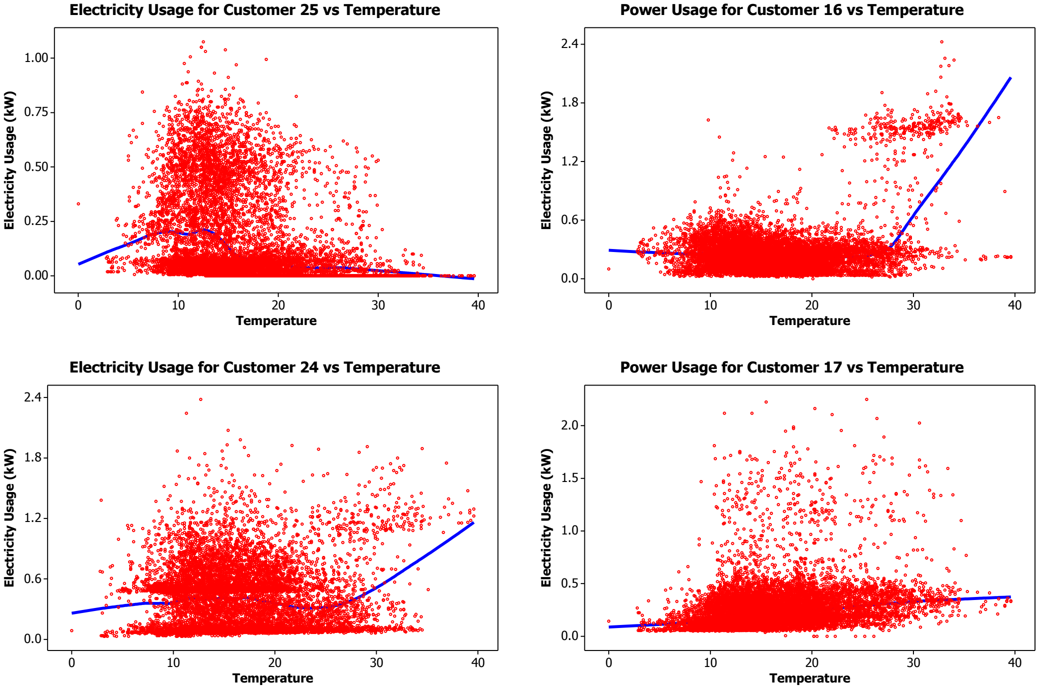

Figure 5: Electricity usage versus temperature for individual customers.

2.3 Power usage by temperature

Figure 5 shows scatterplots of electricity use against temperature for four individual consumers. The blue lines are LOWESS smoothers. The graphs seem to suggest four different patterns of relationship with temperature:

- Customer 25 uses electric heating in low temperatures but no cooling;
- Customer 16 does not use electric heating but uses air-conditioning when it is hot;
- Customer 24 uses both electrical heating and air-conditioning;
- Customer 17 uses neither.

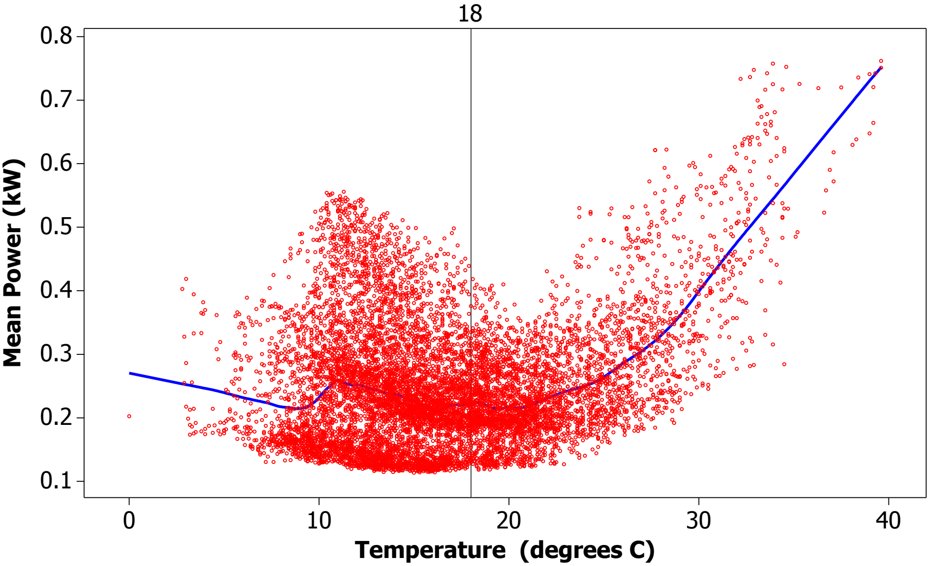

Figure 6: Scatter plot of mean electricity usage across all customers, versus temperature.

These customers were selected because they demonstrate obvious patterns, but the pattern may not be visually obvious for other customers. In practice, with hundreds of thousands of customers, one cannot rely on humans to look at scatterplots for classifying the customers; mathematical rules are needed and these rules will have to cope with a great deal of variation such as customers using air-conditioning on some hot days and not on others. Figure 6 shows the mean electricity use across all 772 customers, by temperature, with each dot referring to a different half-hour in the 199 days. The graph shows two separate clusters, presumably using or not using electric heating. It also confirms the industry view that heating and cooling use tends to be minimised around 18 degrees Celsius (or maybe up to a couple of degrees higher for the Melbourne data).

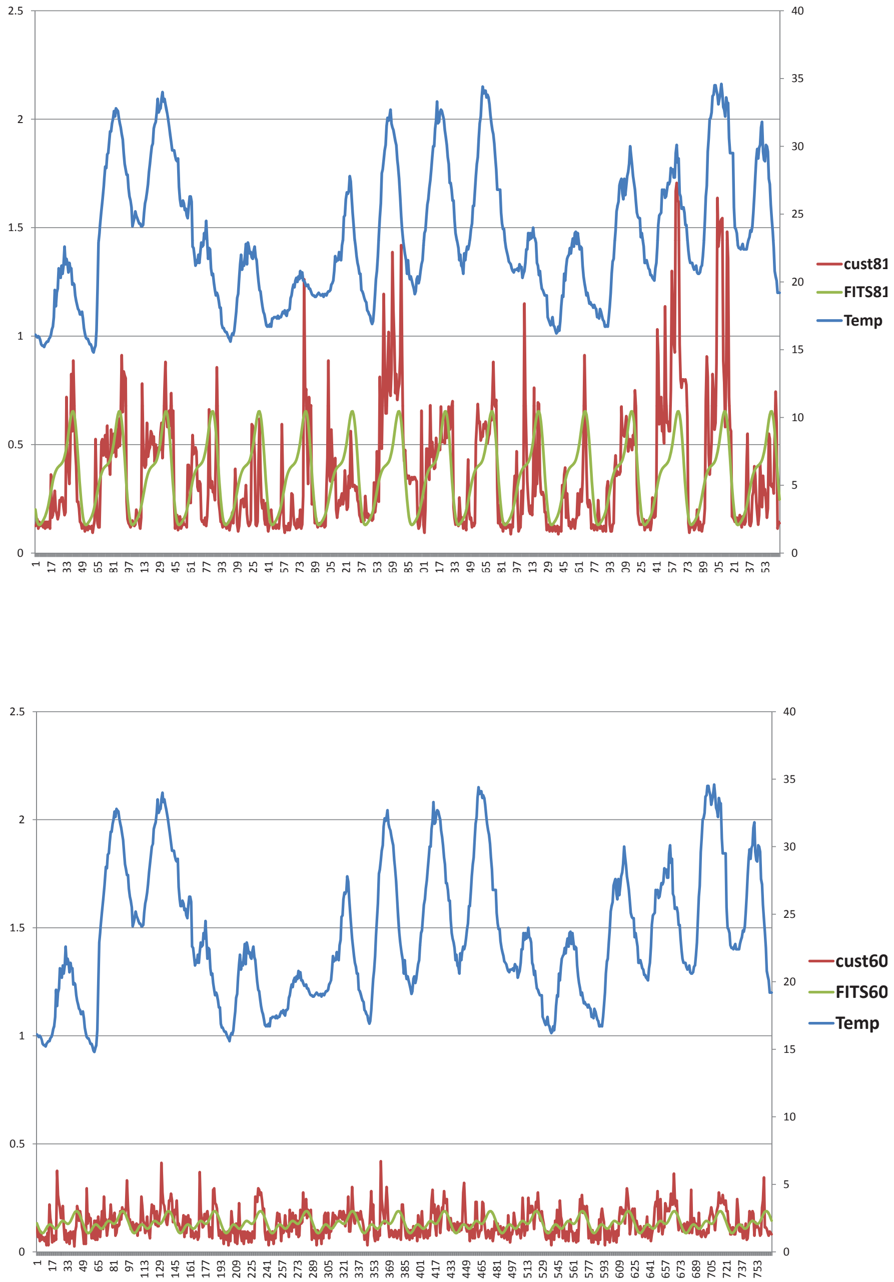

Figure 7: Example of Fourier analysis for two customers.

2.4 Fourier analysis

In theory, customer data could be clustered based on the whole vector of half-hour energy measurements, but this will become unwieldy for very long time series and a large customer database. It therefore becomes essential to summarise the data. One approach used at MISG was spectral analysis of each customer's time series. The rationale was that spectral analysis would look for temporally recurring patterns. Along related lines Figure 7 shows the result of of fitting low dimensional Fourier series models (green curves) to two customers (red curves). The plots represent 16 days, and a rescaled graph of temperature is shown on the same plot. The green curves show one and two peak usage periods per day, respectively, and different magnitudes. It seems likely that a low-dimensional model could adequately express the main differences between customers. That is, the parameters of the Fourier curves may be used as summary statistics for each customer, and these statistics fed into a clustering procedure to divide the customers into groups. Inevitably such a procedure has drawbacks. One drawback is the unrealistic assumption that the time series of electricity use is stationary (an unchanging pattern) and another is that the graphs show the raw data (red curves) vary greatly around the green curves so a lot of information is unrepresented. For example, there is an association between the fitted values and temperature, although temperature by itself does not appear to explain a big proportion of the temporal variation. It might be best to begin by fitting simple linear regression models to express known covariates such as temperature, day of the week, and month of the year (the latter impacting on hours of daylight, holidays, etc.), and then fit a time series model to the residuals. The issue of dealing with gross changes in customer behaviour—as in Figure 3(b)—remains.

2.5 Exploration using Load Duration plots

During the data exploration phase we constructed Load Duration plots for each customer. Figure 8 shows Load Duration plots for three customers. Each plot has three Load Duration curves, for hot (red), moderate (green) and cold (blue) weather. The horizontal axis is the proportion of time and the vertical axis is power. A point (t, p) on a curve indicates that for proportion of time t, the power is greater than p. This form of graph is common in the electricity industry, and is equivalent to a cumulative distribution function for power. The first plot is a Load Duration plot for a customer whose curves are independent of temperature. The second plot is for a customer whose load increases markedly when the temperature is high; the third plot is for a customer who has almost no load, except when the temperature is low.

In principle customers could be classified based on the shapes of their Load Duration curves. For example one could stratify the data by weekend/weekday, time of day (say day 07:00–17:00, evening 17:00–23:00, and night 23:00–07:00) and, say, three temperature ranges. This would give 18 Load Duration curves per customer. The gross features of each Load Duration curve could be summarised by a linear model for log(Load) versus log(Duration), yielding two parameters per curve. Thus the customer's power usage distribution would be summarised by 36 numbers, representing a considerable reduction from 9552 time points. These 36 numbers per customer could be used as the input for a cluster analysis.

The next section considers in detail one method of cluster analysis for customer segmentation. The method was applied to time interval means, but could just as easily have been applied to summary statistics from Load Duration curves, Fourier analysis or other data reduction techniques.

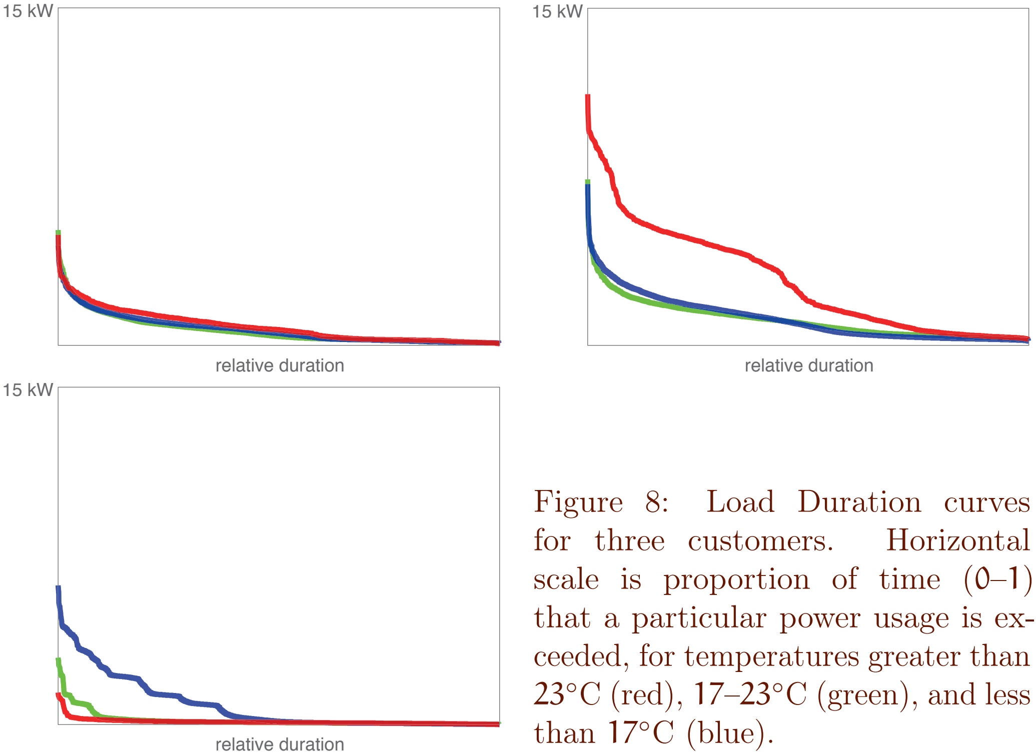

Figure 8: Load Duration curves for three customers. Horizontal scale is proportion of time (0–1) that a particular power usage is exceeded, for temperatures greater than 23°C (red), 17–23°C (green), and less than 17°C (blue).

3 Customer segmentation using Self-Organizing Maps

This section describes the use of Self-Organizing Maps to explore the second question in the project: *Could we identify which customers have significant cooling loads and which customers have significant heating loads?* For each of the 772 customers there were 9552 readings recorded by the Smart Meter. This is a high-dimensional dataset. When the number of variables involved is large, it is normal to consider dimension reduction methods to reduce the dimensionality to a manageable size before applying clustering algorithms for segmentation. One such clustering algorithm is Kohonen's Self-Organizing

Map (SOM) procedure [8]. This procedure maps an n-dimensional input space to a two-dimensional region while assigning each data instance to one of the resulting clusters based on comparing its similarity to all the other data instances. SOM was employed to analyse the mean power use per half hour of the day ($n = 48$ datapoints per customer).

3.1 Self-Organizing Mapping

SOM tries to maintain the original topological structure of the data in terms of some specific similarity measure. That is, data instances that were close in the higher-dimensional input space should remain close in the reduced lower-dimensional map. Given a dataset of n-dimensional vectors v_i, $i = 1, 2, \ldots, m$ (here $m = 772$ customers) this is a dimensional reduction from n dimensions to two dimensions. With respect to a pre-defined similarity measure, the goal is to reproduce in two dimensions the similarity between each pair of data in the given set of n dimensional vectors. The actual dimension reduction procedure involves the following steps:

Step 1 The data input consists of a set of n dimensional vectors v_i, $i = 1, 2, \ldots, m$, standardised to unit scale and with the number m of vectors satisfying $m \gg n$.

Step 2 With respect to the similarity measure to be used, the similarity values S_{ij} are computed for all the pairs v_i and v_j. The similarity measure is often defined in terms of the Euclidean distance between the vectors v_i and v_j. The entries in S_{ij} will be small when v_i and v_j are close. It is these small values which are used as the seeds for performing the subsequent clustering.

Step 3 An arbitrary vector (x_1, y_1) on a two-dimensional square (such as the positive unit square) is chosen to represent the location defined by the vector v_1 in n dimensions.

Step 4 Relative to the position in two dimensions chosen for v_1, an identification procedure is then used to assign, on the basis of the similarity values S_{ij}, the vector (x_i, y_i) in the square that represent the positions of the vector v_i; $i = 2, 3, \ldots, m$, in n dimensions. This is an invariant procedure, in that the relative positioning of the v_j in two dimensions has a similar pattern independent of the choice of the initial point (x_1, y_1).

The clustering phase of the SOM is performed on the two-dimensional data. Its role is the segmentation of the data into similarity groups. Software is available for the application of Ward's procedure [9]. The basic steps are the following.

Step 5a The clustering starts by first identifying some small set of the i's and j's for which the values in the similarity matrix S_{ij} are the smallest, and define the corresponding two-dimensional points (i, j) as the nodes around which the clustering will be performed.

Step 5b Around each node identified at Step 5a, the closest points are collected together to form its cluster.

Step 5c Boundaries are placed around the clusters so that, relative to each node, there is no point in the cluster which is closer to some other node belonging to another cluster.

Step 6 This regional pattern can be utilised in a number of ways to illustrate various relationships within the data. The possibilities implemented include:

- One can overplot, on the resulting two-dimensional regional pattern, some other property associated with the vectors v_i, $i = 1, 2, \ldots, m$, by colour mapping it between the minimum and maximum values of the property;

- When one of the components in the vectors v_i, $i = 1, 2, \ldots, m$, is categorical with K levels then it is appropriate to check whether the realisation generated with K nodes correlates with the categorical variable.

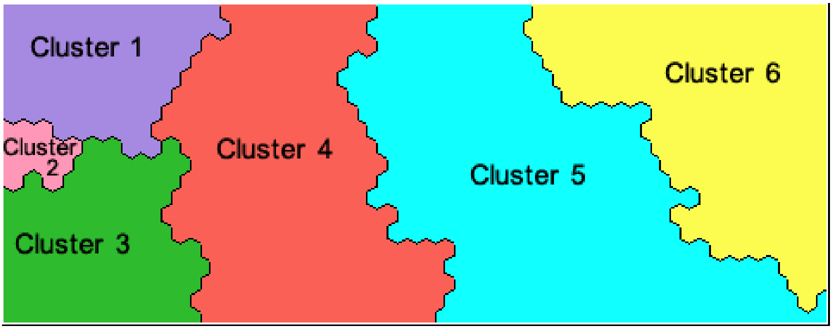

Figure 9: Customer segmentation using SOM.

3.2 Pattern discovered from customer segmentation

As an exploration of this technique, we consider applying SOM to the customers' mean half-hour energy use, averaged across all weeks at that time of day. This still represents 48 data points per individual. Six clusters were obtained, as shown in Figure 9. Random colors were selected to indicate individual clusters. The number of the data instances in a cluster determines the size of each coloured region in the map. That is, a big cluster with more data instances has a larger region in the map. Cluster 5 (light blue) is the biggest cluster obtained by SOM with 246 customers, and Cluster 2, shown in the smallest pink region, has the least number, with just ten customers. Cluster 6 has 215 customers, which is a little bigger than Cluster 4 (166 customers), followed by Cluster 1 and Cluster 3 with similar sizes of 63 and 74.

Each cluster indicates a certain group of customers, whose energy use data may have a different distribution. Figure 10 presents the average energy usage based on half-hour readings for all six clusters, from which a set of patterns was discovered. Based on these patterns, we group the customers into the following four types.

Business customers (Cluster 2) had high energy use between 07:30 and 18:00 and low overnight use between 21:00 and 06:00 the next day.

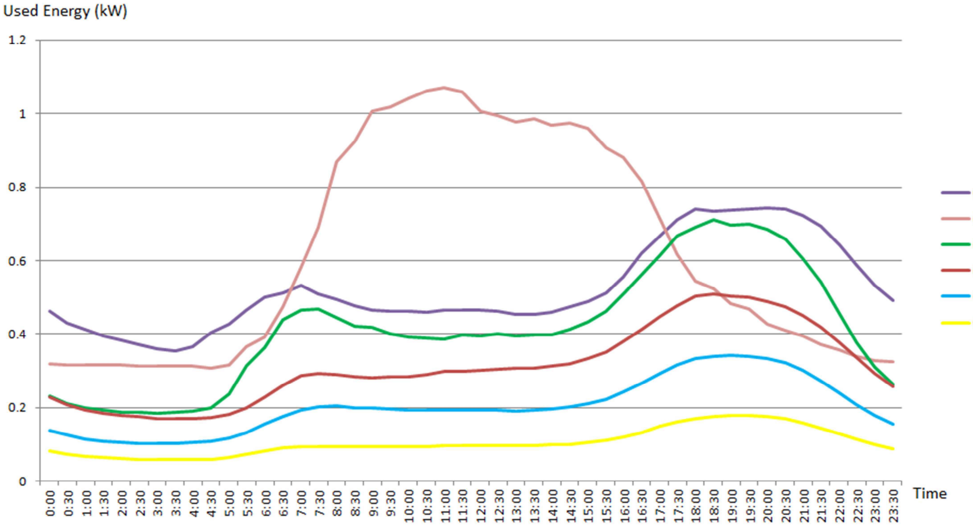

Figure 10: Summary of half-hour energy use for six customer clusters.

High-use households (Clusters 1 and 3) use more energy than the remaining clusters, and in particular use more outside business hours than during business hours. These may be large families, or households that use a lot of heating or air-conditioning. They make up around 18% of the sample. The graph suggests family members got up and had breakfast between 05:00 and 09:00, as indicated by the first peak in Figure 10. After that, they went to school or work in the daytime so that the energy use decreased to a lower level. Then they came back home for dinner preparation and entertainment from 16:00 until bed time, with the energy use reaching its second peak around 18:30 (6:30 pm). It appears the customers in Cluster 1 stayed up later than those in Cluster 3, and also kept more equipment on at night.

Low-use households (Clusters 4 and 5) had similar curves to the high-use households (Clusters 1 and 3) but used less energy overall. Another significant difference was that the energy use did not fall down in daytime. This suggests that these households have someone at home during the day, doing housekeeping, child-minding or working from home. From the two curves in Figure 10, the families in Cluster 4 used more energy than those in Cluster 5, but it is hard to distinguish them otherwise. It might be that Cluster 4 households had more people living in them. These low-use clusters together represent 53% of the sample.

Minimal users (Cluster 6) used less energy than other customers. They might be people with gas heating and cooking facilities, and limited electrical equipment at home, which is used infrequently. It may include single people who are rarely at home. The minimal users make up about 28% of the sample.

The preceding clusters were formed without taking week and temperature into account. Nevertheless, it is of interest to know whether these clusters can tell us whether its customers have electric cooling or heating systems. Figure 11 shows the average daily energy use for the 199 days from July 2011 to January 2012, for each of the six clusters. The time scale ranges from winter, through spring, and into summer, and all clusters show a drop in consumption in the second week of January, probably because of moderate temperatures that week, coupled with people being on holiday.

Customers with electric heating and cooling. Cluster 2, business customers, shows a slight negative trend on the left and possibly a weak positive trend on the right, suggesting that they have both heating and cooling systems at their premises. Their daily energy use stayed at much same level during the entire period except some drops on weekends and between Christmas and 14 January. Cluster 1, on the other hand, indicates families having both electric heating and cooling systems. Their daily energy use was high in winter, with a strong decreasing trend until November when the weather warmed, which indicates they used electric heating. In summer there are several peaks on the same dates that all other clusters used more energy than usual, probably very hot days. The magnitudes of the peaks suggest use of air-conditioning.

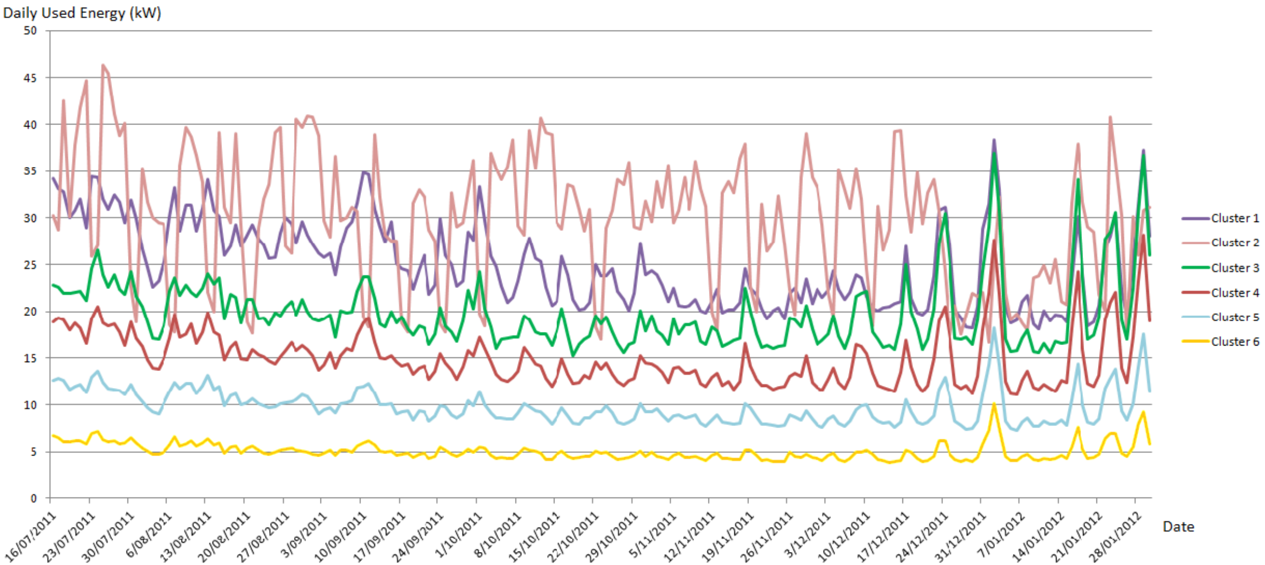

Figure 11: Comparison of average daily energy use for six customer clusters.

Customers with electric cooling only. The customers in Clusters 3 and 4 used cooling systems on the warm days in summer. There were five high peaks from mid-December 2011 to the end of January 2012 as shown in Figure 11. The highest peak reached 38 kWh per day, which was about twice the normal daily use. By contrast, in the cooler months the peaks were less prominent and the downward trend seems weaker than in Cluster 1. The trend may be explained simply by the need for more lighting in winter. Some of these households may use gas heating.

Customers without electric heating or cooling. The daily energy use of Cluster 5 and Cluster 6 do not shown big changes during the entire period. Their energy use stayed much lower than the other four clusters, which showed no high-energy electric heating or cooling systems were used by the customers in these two clusters. They did use more energy than usual in a few summer days, but this may be just due to having fans or portable cooling machines rather than permanent air-conditioners.

4 Clustering based on mean power statistics

A common feature of all the above methods is the need to reduce the long time series of power readings to a lower-dimensional set of summary statistics. In the case of the Fourier methods, the approach was to cluster based on Fourier coefficients. For SOM, the data were first reduced to the average power consumption per half-hour of the day, and these 48 dimensions were further reduced to two dimensions. To allow for more explanatory factors, it might make sense to summarise a customer by their mean electricity consumption in various combinations of circumstances. An example of this approach is described by Chicco [2], who raises the following suggestions. (Some comments are added in parenthesis).

1. Subdivide the data by weekday/weekend and season, with two to three typical weeks of data for each group. (Looking at fewer weeks reduces

the computational load, and also means unusual weeks such as those containing public holidays can be avoided.)

2. Normalise the load pattern by dividing by the typical peak power. (This means that one is looking at what we might call the 'profile' or shape of load pattern, rather than amount of power drawn.)
3. Average over significant periods: night, sunrise, morning, lunch, evening.
4. Calculate 'shape factors' based on ratios (e.g., morning/night), or by Fourier analysis.
5. Group consumers using a variety of possible cluster techniques. (Chicco mentions Adaptive Vector Quantisation, C5.0, Entropy-based, Follow-the-Leader, Fuzzy Logic, Fuzzy and ARIMA, Fuzzy K-means, Hierarchical Clustering, Iterative Refinement Clustering, K-means, Min-Max Neuro-Fuzzy, Multivariate statistics (MANOVA), Probabilistic Neural Network (PNN), Self Organising Maps, Support Vector Clustering, and Weighted Evidence Accumulation Clustering).

The profile idea seems to be an attractive strategy but it does have downsides. It is, after all, reasonable to separate customers out by the actual amount of electricity used, not just by the profile shape. Clustering by shape alone could put customers with very different numbers of electrical appliances (and hence very different total power use) together in the same cluster. Perhaps it would make sense to first stratify the customers by total power use (this can be done without resorting to cluster analysis) and then use clustering techniques separately on each stratum to pick up shape differences. Another drawback of profile analysis is that we would be standardising by typical peak load, which is a random extreme quantity even if based on two to three weeks' data: dividing by it may add noise (redundant variation) to the data. This can make it harder to discern patterns. In particular, for low-use customers their electricity profile may depend on relatively random events (such as the time when the refrigerator motor starts pumping refrigerant). In summary, clustering by profile may help, but is not guaranteed to improve the signal-to-noise ratio of the data.

4.1 Choosing summary statistics for AGL data

At MISG we tried an approach based on summary statistics (means) across various factors. It was assumed that the energy use of each customer depends primarily on time of week and on temperature. At MISG, time was divided into 84 two-hour periods each week, and temperature was divided into five bins. Thus the energy use of a customer was represented by the total energy used in each of up to 420 bins (period, temperature). Some bins may be unrepresented (for example high temperatures in the early morning) but this does not prevent the analysis. The time subdivision accounted for differences between weekends and weekdays, but not for public holidays, which were ignored at MISG. Seasonal variation in energy use was assumed to be just due to temperature (although there is some reason to doubt this assumption due to the effect of lengthening days, daylight saving, and the confounding effect of customers going on holiday). At MISG, customers were clustered using an agglomerative clustering algorithm in Minitab, based on Ward's linkage with Euclidean distance. The results were very similar to the SOM results shown in Figure 10. Following MISG, some modifications were introduced to the method, which seem to have produced better results. In what follows we discuss this modified approach.

At MISG, all customers were treated equally, but it subsequently became clearer that some customers had long stretches of zero power use. This may have occurred when the power was disconnected between different tenancies of a dwelling, or when the homeowner switched off the power for some other reason such as extended travel. In any case these periods start and stop unpredictably, and would affect the results. It was decided, quite arbitrarily, to exclude those customers with more than 25% of half-hour periods with zero power use. This eliminated nine (1% of) customers. For a larger dataset, one might be more stringent, and also might devise ways of classifying the excluded customers, or of matching them to one of the clusters found from the remaining data.

The second modification was to eliminate those days which were public holidays (Melbourne Cup Day, 1 November 2011; Australia Day, 26 January) plus the period from Christmas Day to 2 January (a public holiday that year, in lieu of New Years Day that fell on a Sunday). These days were excluded from clustering because the power usage might be unusual and distort the typical pattern for that time of the week and temperature. Thirdly, the time of day was adjusted for daylight saving, moving time forward by one hour from 02:00 on 2nd October. Correcting for daylight saving meant that the morning and evening peaks were less spread out when averaged over all weeks.

Time of day was coded into 12 two-hour blocks: 00:31–02:30, 02:31–04:30, . . . , 22:31–00:30. This choice of time blocks seemed to correspond best to the weekday circadian pattern. Figure 12 shows the relationship between mean power usage (across all customers) and temperature is fairly similar among each of the four half-hours making up a two-hour block, and yet the pattern is different between two-hour blocks. Temperature was coded into seven intervals: < 10 degrees, 10–13.9, 14–17.9, 18–21.9, 22–25.9, 26–29.9, $\geqslant 30$ degrees Celsius. Days were coded into Weekdays (Monday–Friday) and Weekends (Saturday–Sunday). For each customer, then, their mean electricity use was calculated for each of these $12 \times 7 \times 2 = 168$ combinations, if possible. In practice only 148 combinations were available: that is, there were some combinations of time period × temperature range × weekday/weekend which did not occur for any customers. These unavailable combinations were very low temperatures in mid-afternoon or very high temperatures at night. There was also a handful of customers who had individual missing data means on weekends. Since clustering algorithms require complete data, the corresponding weekday mean was substituted for the missing weekend value. The alternative would have been to drop these customers from the analysis.

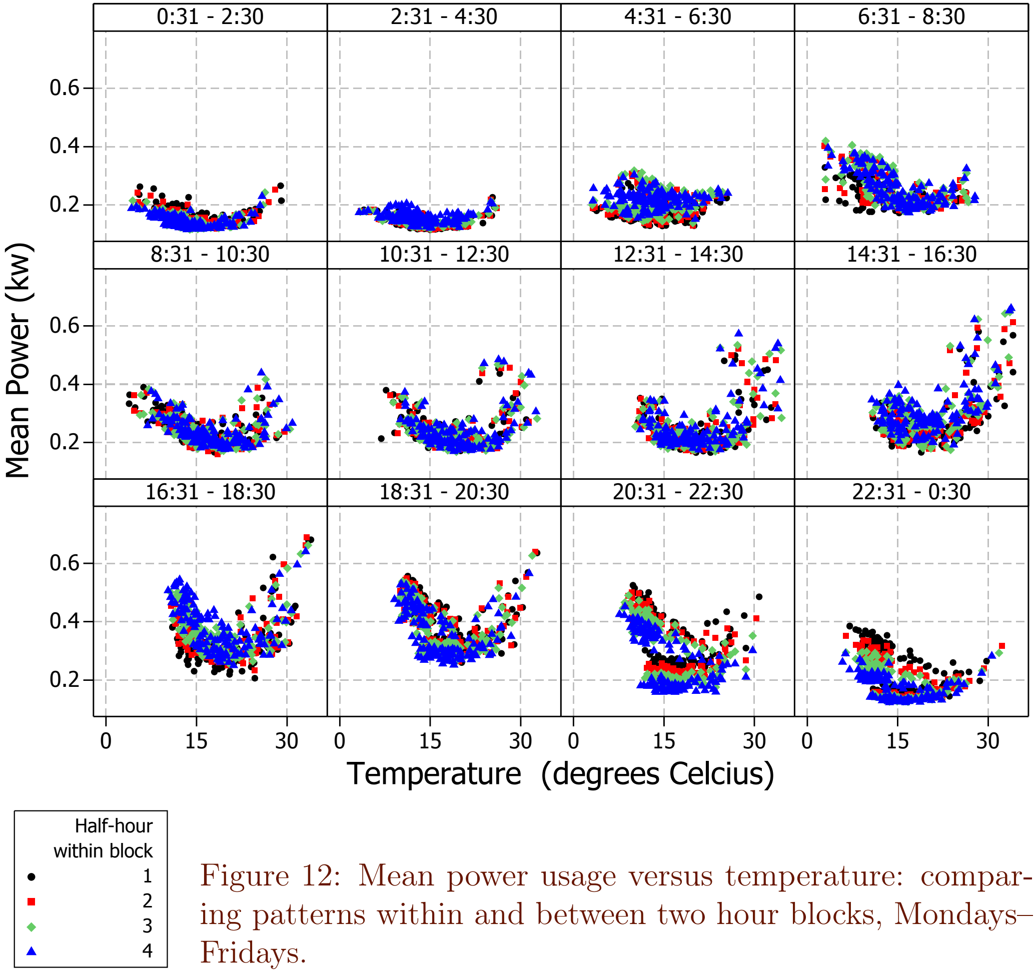

Figure 12: Mean power usage versus temperature: comparing patterns within and between two hour blocks, Mondays–Fridays.

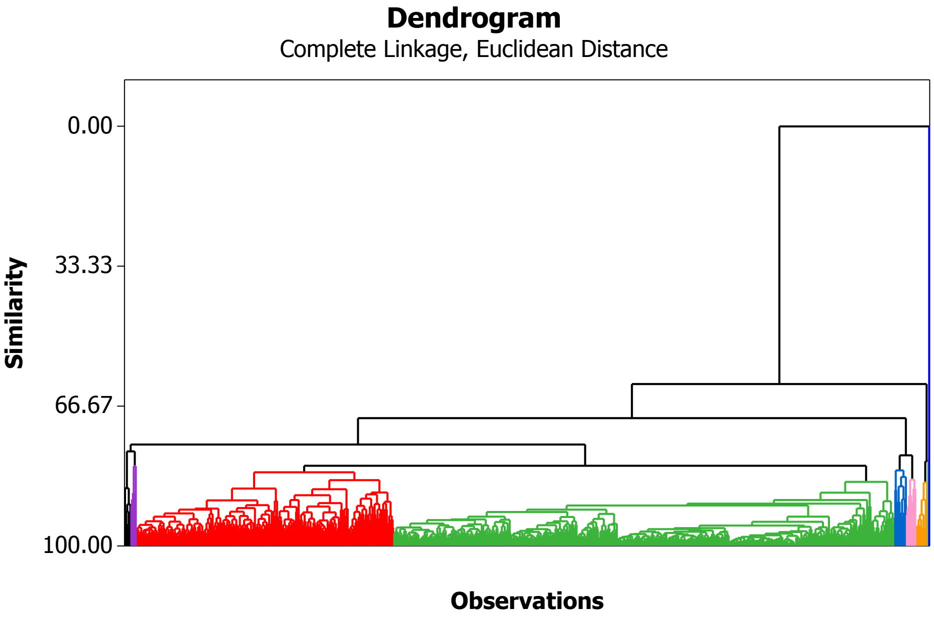

Figure 13: Dendrogram for complete clustering.

4.2 Complete linkage agglomerative clustering with Euclidean distance

To begin clustering we must decide on a measure of distance between any two individuals. Here we focus on Euclidean distance; that is, for every available combination of time period × temperature range × weekday / weekend, the difference is calculated between the mean power usage by customer A and by customer B, then this difference is squared, added over all available combinations, and then the square root taken. The effect of squaring is that the bigger differences dominate; that is, one large difference is more important than many small differences. This is in contrast to, say, the Manhattan distance measure which is based on minimising the sum of absolute differences, and so gives relatively more importance to many small differences in power usage compared to one bigger difference. At the other extreme, squared Euclidean difference gives even more emphasis to the biggest single discrepancy between customers A and B.

Most clustering routines are agglomerative; that is, they begin by assuming every customer is in its own little cluster, and then clusters are joined together, starting with the clusters that are the smallest distance apart. Step by step, more and more clusters are joined until finally the whole dataset is joined together as a single cluster. The process of building up and linking clusters is illustrated by the Dendrogram in Figure 13. Clusters that join at the top of the graph are the most different in the dataset—relatively speaking their similarity is zero. Clusters that join near the bottom of the graph are relatively close together by Euclidean distance; their relative similarly is nearly 100%. Typically one chooses a small number of clusters; here nine are shown, of which two are singletons—customers who are very unlike any others. The singleton on the extreme right of the graph is the high electricity-use business in Figure 2(b). The second singleton (hard to see on the graph) was the one shown in Figure 3(a). In principle one chooses the number of clusters by picking a level of similarity, drawing a horizontal line, and separating off those clusters cut by the horizontal line. In practise one works in reverse, choosing the number of clusters required and then the algorithm draws the line accordingly. The graph shows seven other clusters (besides the two singletons) represented by different colours in the dendrogram. This is the maximum number of clusters we can find from the graph without splitting off many more singletons, which would seem to defeat the purpose of clustering.

The other key aspect of an agglomerative clustering algorithm is the linkage rule. The linkage defines conditions for the combining of two clusters, and is of prime importance in defining the shape of the dendrogram. Complete linkage, used here, is a rule that says that *every* pair of customers within a given cluster must have at least the required level of similarity (that is, be no more than a certain maximum distance apart). This means that if unusual customers are already within a cluster, it is hard to find other customers to join them. So complete clustering tends to give quite distinctive clusters,

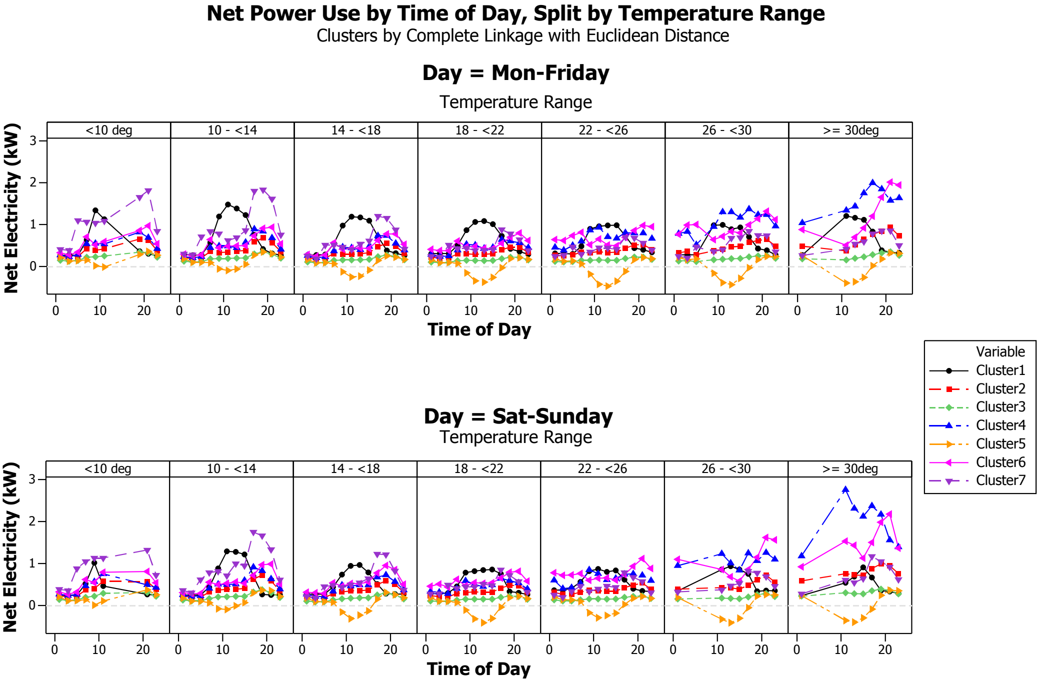

Figure 14: Profiles for clusters defined using complete linkage with Euclidean distance: showing net electricity use by time of day, split by type of day and by temperature range (degrees).

which may vary greatly in size. In Figure 13 most clusters are small but 62% of customers belong to one cluster (green) and another 32% belong to a second cluster (red). If we reduced the total number of clusters by one, then these two clusters would merge while the smaller clusters would stay distinct.

Figure 14 depicts the profile of these clusters in terms of average power use (cluster centroid) by time of day, when the temperature is within various ranges, for weekdays and weekends. Cluster colours match those of the dendrogram. In addition to the two singletons mentioned earlier (not graphed), the clustering procedure has identified a group of six other customers (Cluster 1, black dots) who appear to be small businesses with high electricity use during the day and little at night, and less prominent energy use on weekends than Monday–Friday. Complete linkage clustering has also identified a group of eleven customers (Cluster 5, orange triangles) with inverted net use profiles. These must be the PV customers who at certain times export electricity back to the grid, and so their net use can go negative. As might be expected for solar power, the negative profiles are fairly symmetrical around midday, with depth increasing as temperature (and by implication, sunlight) increases, and the profiles are fairly independent of the day of the week. They are joined by the singleton from Figure 3(a) further up the dendrogram. Cluster 7 (inverted dark purple triangles) is a small group of only five customers who have very high electricity usage when it is cold but moderate when it is warm (over 18 degrees). Clusters 4 (light blue triangles, eleven customers) and 6 (pink triangles, nine customers) both use little electricity when it is cold or moderate, but very high amounts when the temperature gets into the range above 22 degrees Celsius. These clusters are divided from the same branch of the dendrogram, and could be considered one group, but they are distinguished by the amount of electricity drawn, and by higher weekday afternoon use of air-conditioning in Cluster 4. Presumably Cluster 4 customers stay home and use the air-conditioning while Cluster 6 customers go to work and turn the air-conditioner on when they get home. Perhaps these are the type of customers that could be incentivised to have photovoltaic cells installed. The largest group of customers (Cluster 3, green diamonds, 475 customers) are ones that on average use very little electricity at any time, and the second-biggest group (Cluster 2, red squares, 244 customers) also have a low, fairly constant, electricity profile, but use a bit more when it is cold or hot. In conclusion, complete clustering with Euclidean distance appears to be adept at distinguishing small meaningful clusters.

It is of interest to know what might be the effect of using Manhattan distance instead of Euclidean distance. With this measure the two singletons mentioned above were identified, plus one other singleton and a cluster of size two. It is harder to interpret the results if there is a proliferation of singletons or small clusters. The extra singleton (not graphed) was a customer with a large negative net electricity use in the middle of the day. The electricity use profile for the other clusters is shown in Figure 15. The Cluster 8 (pink +) is the cluster of two individuals who have very high electricity use at breakfast time and evenings, especially when it is cold. The small cluster of business customers (Cluster 1, black dots) is also identified by Manhattan distance. The four remaining clusters, which make up 98.5% of the customers, do not show much difference in *shape* with time of day; the profiles run largely in parallel within each panel of the graph. The main differences are with temperature. For Cluster 5 (orange triangles, 38 customers) power consumption is high in cold weather and low when it is hot; for Cluster 4 (blue triangles, 49 customers) power consumption is fairly constant in cool weather but rises steeply in hot weather; and for Cluster 2 (red squares, 202 customers) and Cluster 3 (green diamonds, 462 customers) power consumption is higher when it is either cold or hot, and lower around 18–22 degrees. Clusters 2 and 3 just differ in the amount of energy consumed.

4.3 Complete linkage clustering with Manhattan distance

In summary, the benefit of using Manhattan distance has been to have larger clusters, with a simple interpretation in terms of the relationship between power use and temperature. On the other hand, the routine has not found much shape difference with time of day, between the clusters. A concern is that the only customers identified as going negative were two singletons, as opposed to a group of eleven customers with Euclidean distance. The other negative ones have been hidden by the clustering procedure. Also there are more singletons or tiny clusters, which make interpretation difficult.

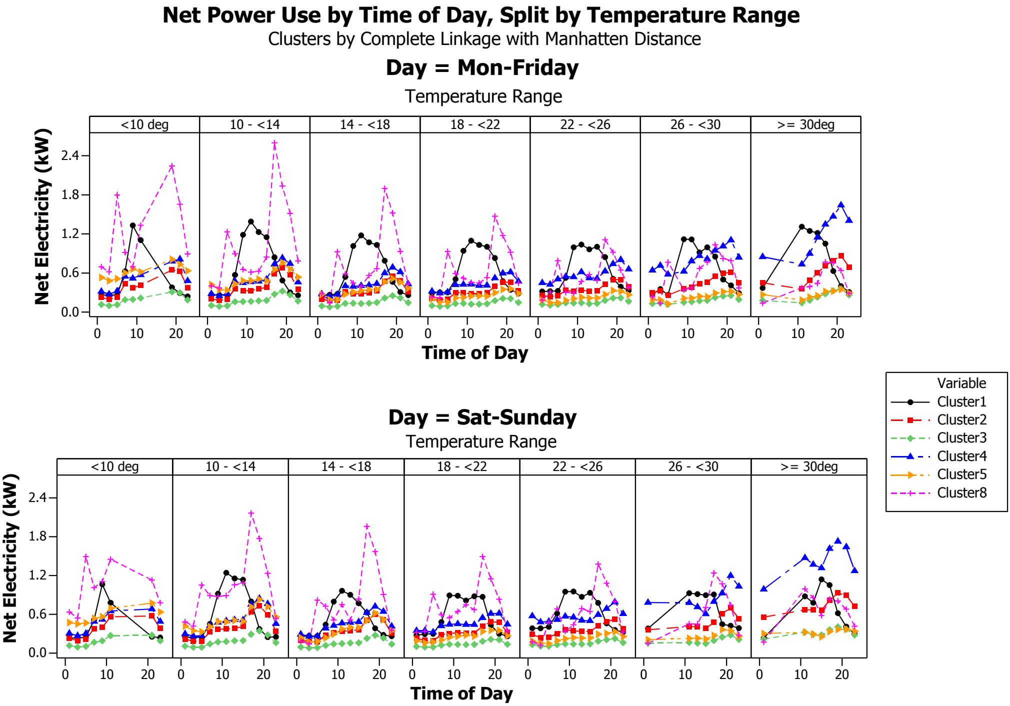

Figure 15: Profiles for clusters defined using complete linkage with Manhattan distance: showing net electricity use by time of day, split by type of day and by temperature range (degrees).

4.4 Ward's linkage clustering with Euclidean distance

Figure 16 shows the dendrogram for Ward's linkage. As mentioned in the online Minitab help information:

> With Ward's linkage, the distance between two clusters is the sum of squared deviations from points to centroids. The objective of Ward's linkage is to minimise the within-cluster sum of squares. It tends to produce clusters with similar numbers of observations, but it is sensitive to outliers.

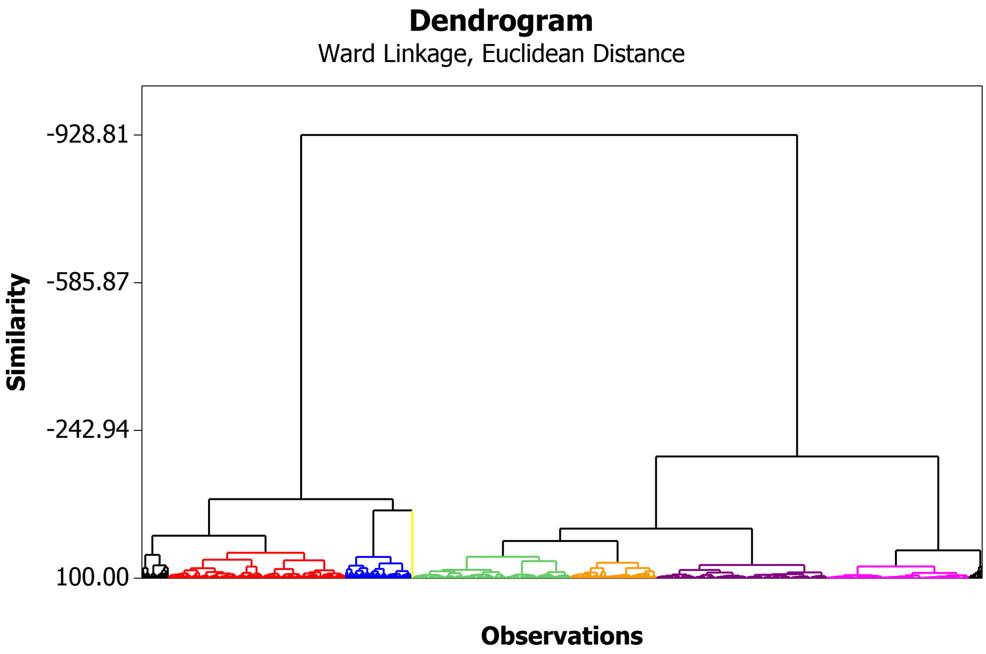

Figure 16: Dendrogram for Ward's Linkage with Euclidean Distance

The nature of the sensitivity is that Ward's linkage tends to find fewer singletons, because they are incorporated into clusters. The advantage is a dendrogram that looks visually simpler, but the disadvantage is that the centroid of the cluster is affected by outliers, which distorts the interpretation of the pattern. In Figure 16 we chose to display eight clusters (excluding the business singleton from Figure 2(b)). The cluster profiles are shown in Figure 17. The method has succeeded in identifying the cluster of customers with PV systems (Cluster 8, black +, twelve customers, extreme right of the dendrogram; the singleton from Figure 3(a) is now included with the other customers with negative usage). However, the cluster of business customers found with complete linkage has now been merged with other high-use domestic customers (Cluster 1, black dots, 24 customers) so the shape profile is now showing an early evening peak when the weather is cold.

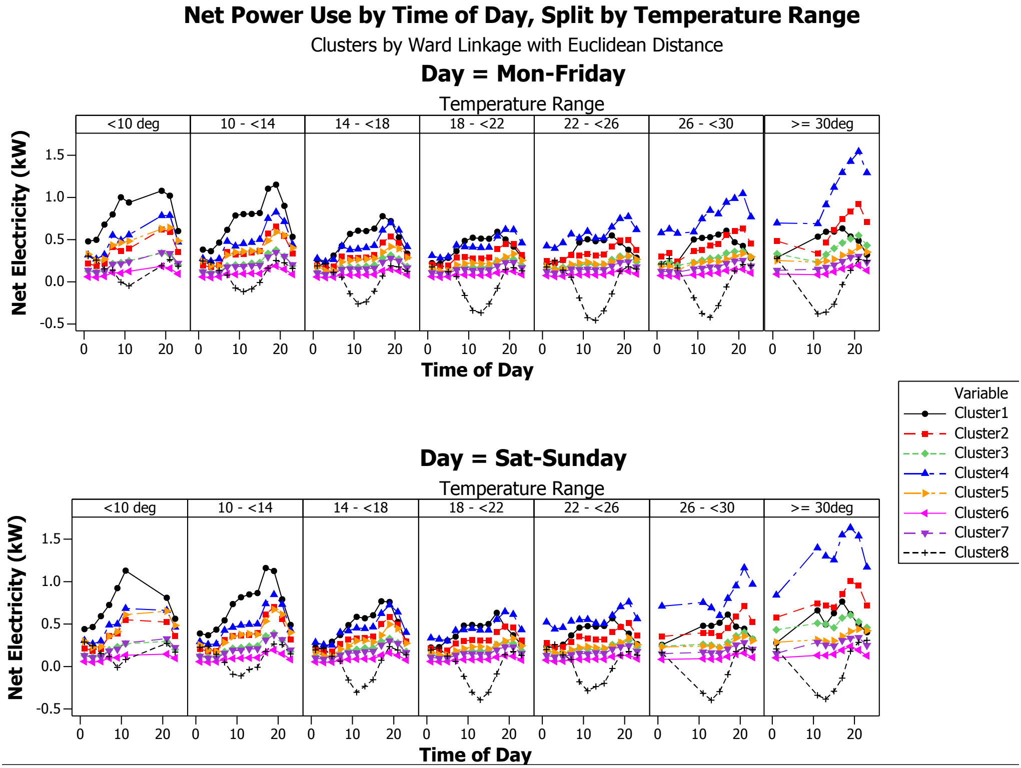

Figure 17: Profiles for clusters defined using Ward's linkage with Euclidean distance: showing net electricity use by time of day, split by type of day and by temperature range (degrees).

If eleven clusters were allowed, then this cluster would separate out into two high-use groups, one being seven businesses and the other being 17 domestic customers with high use of electricity in cold weather (see Clusters 1a and 1b in Figure 18). Cluster 2 (red squares, 161 customers) is a group of customers with more power use in both cold and warm weather. Again, if eleven clusters were allowed, then Cluster 2 would split into two groups which are fairly similar in the daytime but one (85 customers) uses more heating and cooling at night (see Clusters 2a and 2b in Figure 18). Back in Figure 17, Cluster 5 (orange triangles, 78 customers) uses more electricity in cold weather but not when it is hot. Cluster 4 (blue triangles, 60 customers) have high power use all the time but much higher for running air-conditioning in hot weather. Cluster 3 (green diamonds, 143 customers) has low power use most of the time but uses a bit more in hot weather. Cluster 6 (pink triangles, 129 customers) has very low power use at all times and in all weathers, and Cluster 7 (purple upside-down triangles, 155 customers) runs parallel with it, just at slightly higher usage levels.

At MISG results were presented for Ward's linkage with just six clusters excluding the business outlier from Figure 2(b). It is of interest to compare the profiles when there are six clusters to those with eight—see Figure 19. In this case Clusters 6 and 8 merged (graphed using pink double circles), hiding the PV customers amongst the rest of the low-usage group. Also, the previous Cluster 3 (little power use in winter, more in summer) merged with Cluster 5 that had the reverse weather behaviour (more power use in winter, less in summer) to produce a cluster with little to distinguish it from the others except for the moderate magnitude of total power use.

In conclusion, provided one allows for enough clusters, Ward's method does seem to be able to find clusters that have an interpretable profile. The disadvantage of Ward's method (compared to complete linkage) is that we may have to look at a large number of different clusters in order to find those truly unusual groups that complete linkage identified straight off. As well as the extra complexity of having many groups, the question is how reliable some of these Ward clusters really are, since they are based on small differences among small numbers of customers.

4.5 Other agglomerative clustering methods

Average linkage is a method that joins clusters based on the mean distance between observations in one cluster and those in another. Figure 20 shows the dendrogram for average linkage with twelve clusters. However, eight of

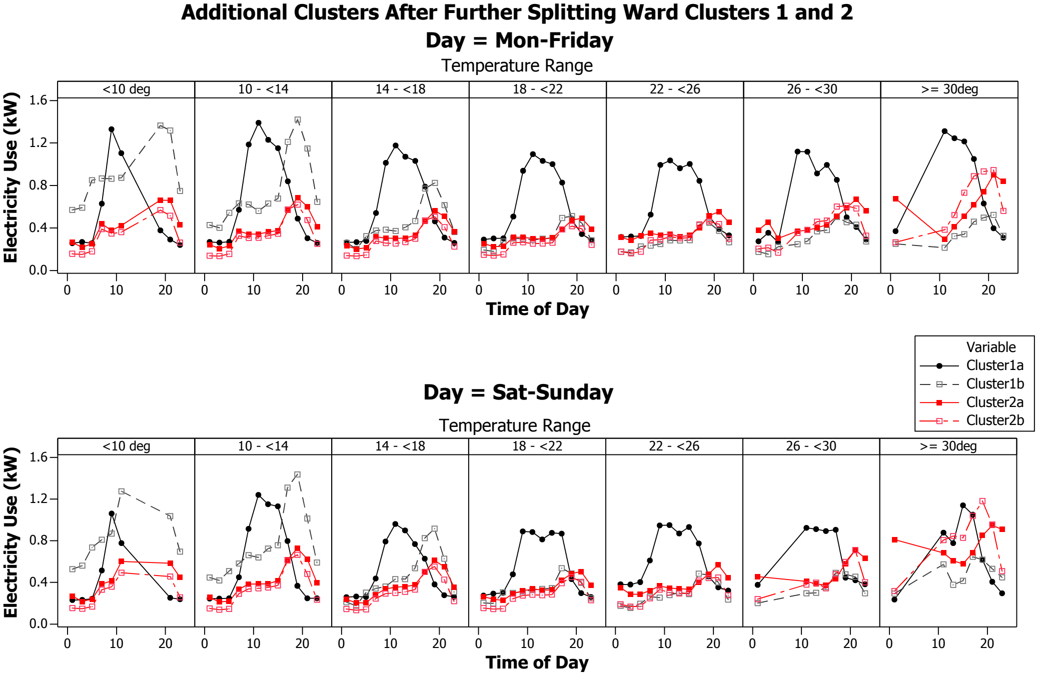

Figure 18: Profiles for additional Ward clusters after further splitting Clusters 1 and 2.

those clusters are singletons, one was a pair, and the remainder comprise five, six and 742 customers respectively. The cluster on the left is our business customers and we needed twelve clusters before they were separated from the rest of the 742. The PV customers are not identified.

Most other methods of clustering (single linkage, centroid linkage, median linkage) fared worse, picking off only singletons one at a time from the main cluster. McQuitty linkage (Figure 21) did produce clusters of meaningful size, but five out of the first eleven clusters were singletons. McQuitty's linkage works by joining the closest clusters, not on the basis of how distant clusters *are* from each other on average, but on the basis of how distant a newly-combined cluster *will be*, from the others, on average. McQuitty linkage

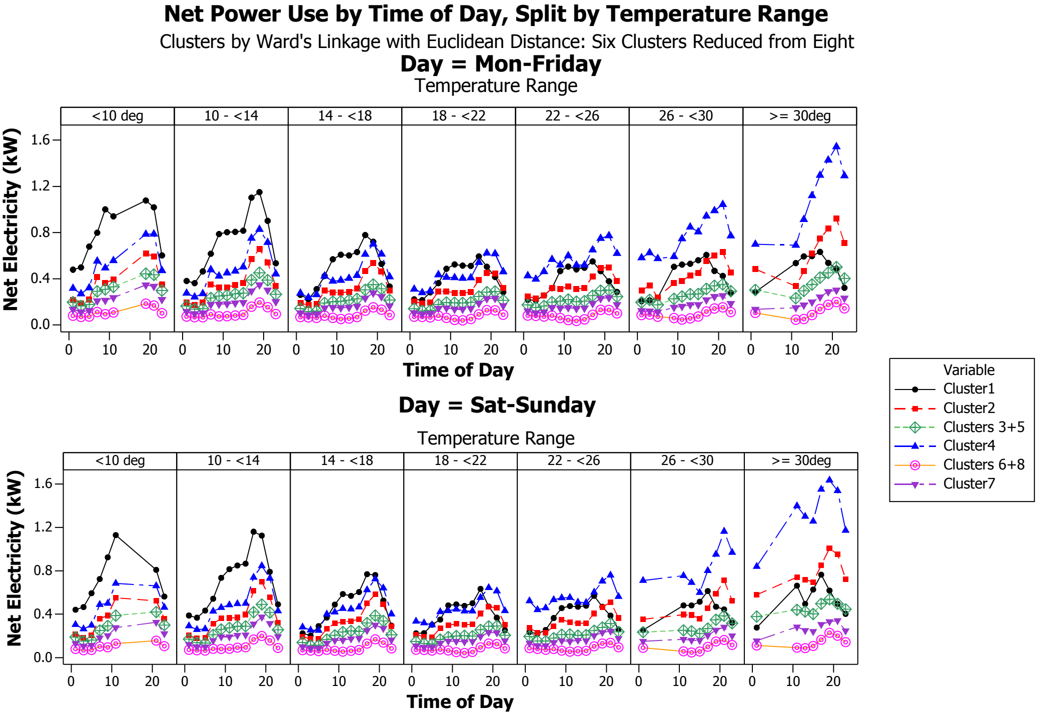

Figure 19: Profiles for Ward's linkage with only six clusters, showing net power use by time of day, split by type of day and temperature range.

found the cluster of small businesses (Cluster 1 in Figure 22) but did not succeed in identifying the PV customers. It also did not identify residential customers with high heating bills in winter, but did find two small clusters with very high electricity use in summer: Cluster 4, 7 customers with high electricity use all day and night when it is hot; and Cluster 6, 10 customers with high electricity use only on hot evenings. Presumably Cluster 6 are out working during the day.

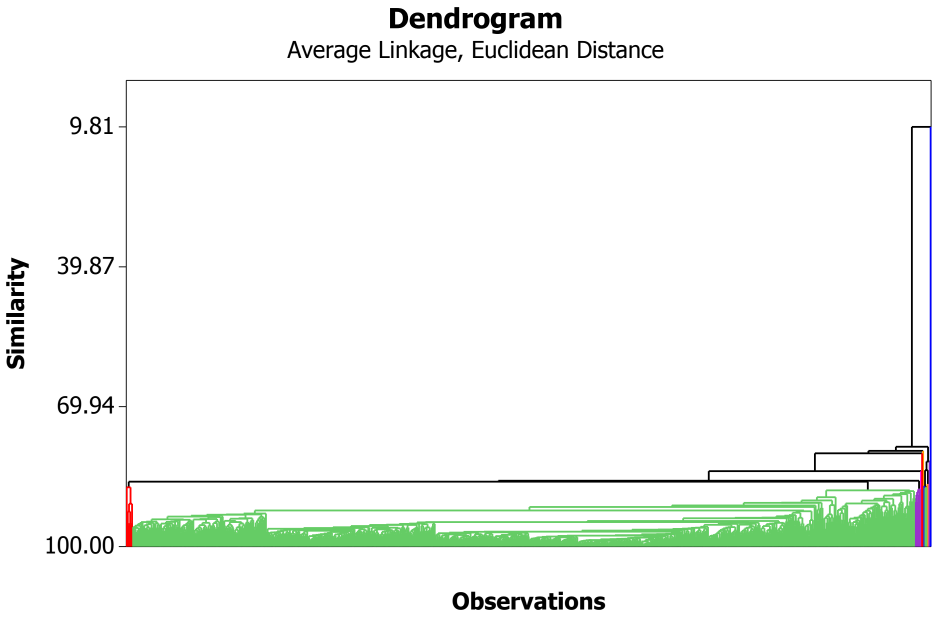

Figure 20: Dendrogram for average linkage with Euclidean distance.

4.6 K-means clustering

The other popular form of clustering is K-means clustering. This is a non-agglomerative procedure which does not produce a dendrogram. One must specify the number of clusters, and preferably estimate initial cluster centroids. If not, then the routine chooses its own starting points for clusters, and then customers are allocated to the cluster with the nearest centroid. The cluster centroid then is recalculated, and customers reallocated to the cluster with the (new) nearest centroid. The process iterates until no changes are made. Figure 23 shows the result of assuming there were twelve clusters. The usual singleton business was one (not shown), and the remaining eleven clusters are graphed in Figure 23. K-means clustering only found the PV customers (blue dots) when there were twelve or more clusters assumed. The procedure also found the cluster of small businesses (yellow triangles), two clusters with

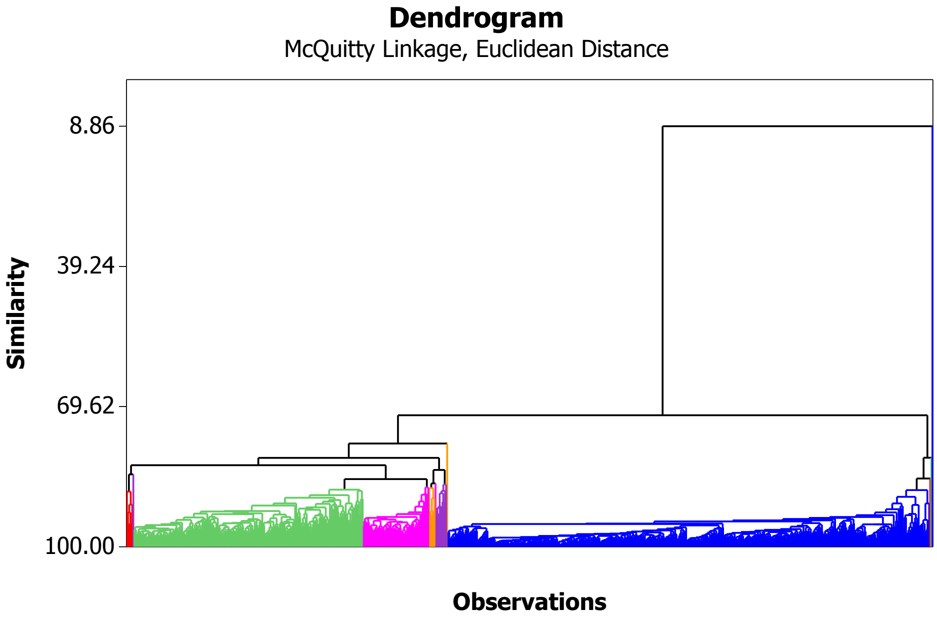

Figure 21: Dendrogram for McQuitty linkage with Euclidean distance.

high power use in summer (green triangles, black +), and one cluster with more power use in cold weather than other weather (purple upside-down triangles). The remaining six clusters all had much the same shape with time and temperature, but differing magnitude of power use.

In conclusion, K-means clustering managed to identify some unusually-shaped clusters but it required assuming a large number of clusters. A disadvantage of K-means clustering is that there is no dendrogram, which means there is no indication how the clusters may relate to each other. On the other hand, K-means clustering has the major advantage that one could use a pilot study to the estimate cluster centroids, and feed those in to K-means clustering to cluster a much bigger population. Furthermore, once the population is clustered, the database could be periodically re-clustered using the latest set of cluster centroids. By contrast, computational constraints mean it would

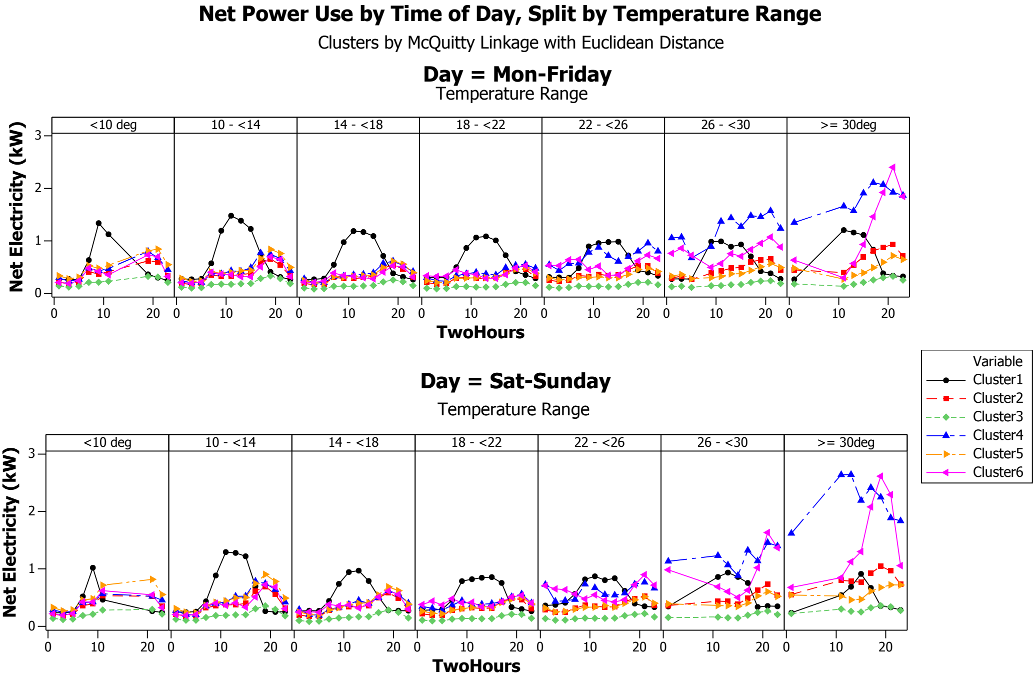

Figure 22: Profiles for clusters defined using McQuitty linkage with Euclidean distance, showing net power usage by time of day, split by type of day and temperature range.

be impossible to use agglomerative clustering for a population of hundreds of thousands of customers, since every cluster (initially, every individual) has to be compared to every other cluster multiple times. Perhaps a way forward is to use an agglomerative routine on a somewhat bigger pilot sample, to find initial clusters with meaningful shape characteristics, and thereafter use them as starting values for K-means clustering.

5 Discussion

There are many methods that can be used to partition customers into clusters. The unusual feature of the AGL task is that each customer is represented by a very long time series of power readings: each sample series was 9552 data points $= 199$ days $\times\, 48$ half-hours, and yet in principle one could use an even longer series, with multiple years of data. This necessitates the use of some sort of data reduction to summarise each customer, before inputing into the clustering routines. The task's complexity is increased by having to deal with negative net power (due to PV systems), zeros (periods when the power was switched off, either because of temporary outage or absence for several days), public holidays or customer holidays when normal patterns of power use are disrupted, and the change to daylight saving time.

We recommend that some protocols for data cleaning should be established before attempting to cluster the customers, to remove business customers and customers with excessive power use or very long absences. Public holidays (or if necessary, the surrounding weeks) should be removed before clustering, and that the time series be adjusted for daylight saving.

More work could be done on the use of Fourier analysis to characterise customer behaviour, but one also needs to adjust for temperature and day of the week. Perhaps a way forward might be to summarise each customer by their regression coefficients from a linear model where their power use is regressed on the temperature (heating degree days and cooling degree days), day of the week, and some sinusoidal curves (perhaps up to order three or four). The coefficients could then be input to a clustering routine. Similarly, more work could be done on Load Duration curves, which could be stratified by time, day, and temperature range, and then summary statistics extracted from each curve. The advantage of modelling Load Duration curves is that more of the distribution of power usages is considered, not just the mean power usage in each stratum.

Notwithstanding the above comments, most attention was paid to summarising

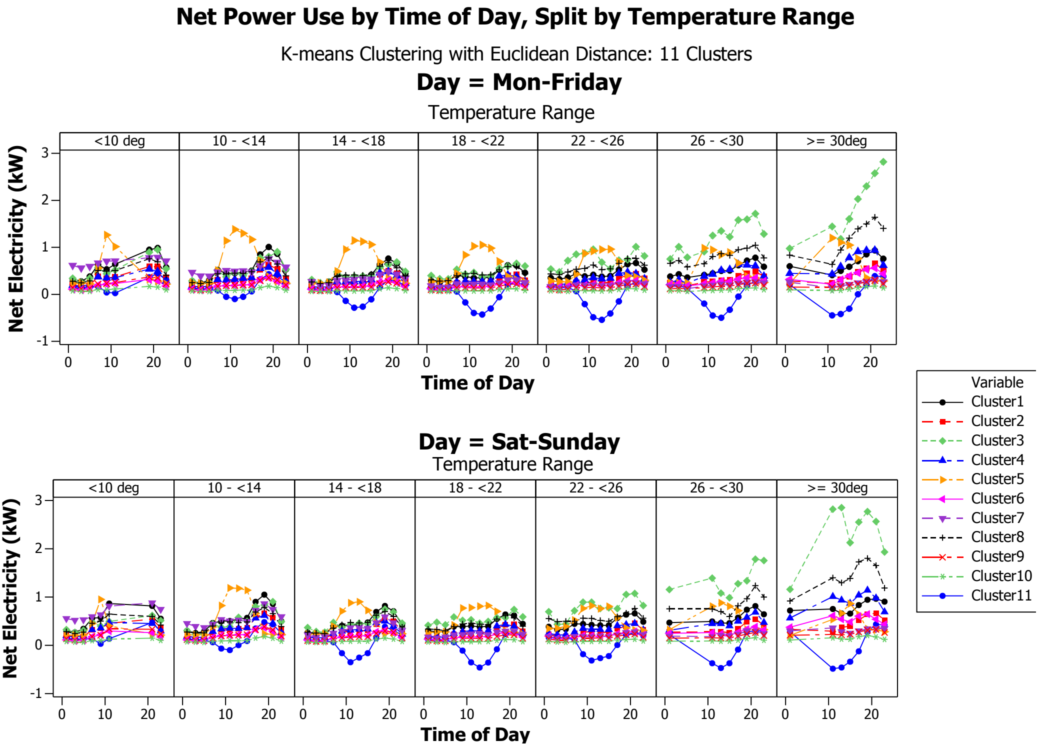

Figure 23: Profiles for clusters defined using K-means clustering with Euclidean distance, showing net power usage by time of day, split by type of day and temperature range.

customers by their mean power use over a period. For input to SOM, customers were summarised by their mean power in each of the 48 half-hour periods; this could be extended to allow for different temperature ranges. In Section 4, customers were summarised by their mean power use in two-hour blocks, in seven temperature ranges and on weekends and weekdays. This still represents a considerable simplification of the data, but gave enough detail to show a lot of variation in electricity use pattern. It also permitted the cluster centroids to be graphed in an interpretable way.

Various clustering techniques were explored in Section 4. Complete clustering seemed to be good for picking out really unusual clusters. Ward's clustering worked well provided one allowed for a large number of initial clusters. In a sense complete linkage, by picking out small but meaningful clusters, provided the criteria for deciding how many clusters are needed for Ward's method. The other agglomerative cluster routines tended to produce far too many singleton clusters (outliers). K-means clustering looks promising for application to large datasets of AGL's size, but needs initial cluster centroids. For future work, we recommended that complete and Ward clustering be used on a larger pilot study, to find interpretable clusters, and then these be used as initial estimates for K-means clustering of the population dataset.

A limitation of Section 4 is that all the clustering routines assume that each data point is of equal worth. However, some of the time × temperature range × weekend/weekday combinations had no data or sometimes only one or two original data points contributing to their mean power. Other combinations had mean power based on over one hundred original data points. This is not *necessarily* a problem—treating all means as having the same worth gives more weight to the customer behaviour in extreme temperatures, and this may help us identify interesting clusters. However, for future work it is worth considering the use of a cluster routine that gives different weights to different data points.

Acknowledgements We thank the industry presenters, Ben Howard and Steve Leopardi for presenting the problem to us, and for their invaluable presence, encouragement and insights. We are very grateful to the MISG participants, who explored the data in a variety of ways, only some of which have been able to be mentioned here: John Boland, Julia Piantadosi, Xuan Vu, Ben Stott, Ryan Steed, Alasdair McAndrew, Shaowu Liu, Tony Gibb, Jason Whyte, Lyndon Walker, Frank Jackson, Josh Chopin, Bogdan Masznicz, Dan Kildea, Rohit Kumar and any others that we may have missed. Finally we thank AGL for agreeing to sponsor this problem and providing the data.

References

[1] Chicco, G. (2009) Support vector clustering of electrical load pattern data. IEEE Transactions on Power Systems, 24: 1619–1628. doi:10.1109/TPWRS.2009.2023009 M107

[2] Chicco, G. (2010) Clustering methods for electrical load pattern classification. In: 8th World Energy System Conference (WESC 2010), Targoviste, Romania, 1–3 July 2010. pp. 5–13 doi:10.1016/j.energy.2011.12.031 M107, M127

[3] De Silva, D., Yu, X., Alahakoon, D., Holmes, D. (2011) Incremental pattern characterization learning and forecasting for electricity consumption using smart meters. In Industrial Electronics (ISIE), 2011 IEEE International Symposium on (pp. 807–812). http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05984262 M107

[4] Flath, C., Nicolay, D., Conte, T., Dinther, C., Filipova-Neumann L. (2012) Cluster Analysis of Smart Metering Data. Business Information Systems Engineering 4: 31–39. http://aisel.aisnet.org/bise/vol4/iss1/5 M107

[5] Gullo, F., Ponti, G., Tagarelli, A., liritano S., Ruffolo, M., Labate D., (2009), Low-voltage electricity customer profiling based on load data clustering, IDEAS 2009, pp. 330–333. doi:10.1145/1620432.1620472 M107

[6] Kim, Y. I., Kang, S. J., Ko, J. M., Choi, S. H. (2011a). A study for clustering method to generate Typical Load Profiles for Smart Grid. Power Electronics and ECCE Asia (ICPE & ECCE), 2011 IEEE 8th International Conference on. pp. 1102–1109. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05984262 M107

[7] Kim, Y. I., Ko, J. M., Choi, S. H. (2011b). Methods for generating TLPs (typical load profiles) for smart grid-based energy programs. In Computational Intelligence Applications In Smart Grid (CIASG), 2011 IEEE Symposium on (pp. 1–6). doi:10.1109/CIASG.2011.5953331 M107

[8] Murtagh, F. (1995) Interpreting the Kohonen Self-organizating feature map using contiguity-constrained clustering. Pattern Recognition Letters, 16:399–408. doi:10.1016/0167-8655(94)00113-H M121

[9] Ward, J. H. Jr (1963) Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58: 236–244. doi:10.1080/01621459.1963.10500845 M122

Author addresses

1. **Barry McDonald**, Institute for Natural and Mathematical Sciences, Massey University, Auckland 0632, New Zealand
`mailto:B.McDonald@massey.ac.nz`

2. **Peter Pudney**, Centre for Industrial and Applied Mathematics, University of South Australia, Mawson Lakes 5095, South Australia
`mailto:Peter.Pudney@unisa.edu.au`

3. **Jia Rong**, Melbourne Institute of Business and Technology, Deakin University, Burwood 3125, Australia
`mailto:Jia.Rong@tulip.org.au`