Mark Goldsworthy

Air-conditioner Use Prediction from Smart Meter Data

Mark Goldsworthy

CSIRO Energy, 10 Murray Dwyer Cr. Mayfield West, Newcastle, Australia
E-mail: mark.goldsworthyr@csiro.au

Abstract

One of the barriers limiting greater uptake of energy efficient consumer behaviors is the lack of information relating to how much electricity is consumed by different devices in the home at different times. Methods of obtaining this information are available but for most consumers the effort and cost involved is unjustified. Ideally the information would be provided freely to consumers, for example as part of their normal energy bill. On average, hot water systems and air-conditioners are the two biggest users of energy in Australian households. This paper described one approach to estimate air-conditioner usage of individual households based on data from their smart meter recorded at 30 minute intervals, combined with weather data. Results show that the addition of information from just a couple of known sampling points, for example as obtained by simply asking the occupant if air-conditioning is in use, greatly improves the overall accuracy of the air-conditioner prediction model for a full year of operation.

1. Introduction

Understanding electricity use is becoming increasingly important given rising electricity prices and greater awareness of the impact of electricity production on the environment. Typically consumers receive information on their consumption in the form of an energy bill that contains aggregate total consumption values, often over a period of several months. Generally there is no detailed breakdown of usage patterns, and almost certainly no information on consumption from specific appliances, with the possible exception of devices connected to controlled load tariffs such as hot water systems.

At the same time, it is recognised that providing consumers with more detailed information is critical to enabling informed choices; for example to drive technology uptake, changing electricity retailers, lifestyle choices or behaviours (Iwayemi & Zhou, 2014). This is most relevant to the appliances that consume the majority of electricity. Air-conditioners, and hot water systems are estimated to account for 41% and 18% of residential energy use respectively (DEWHA, 2008).

Separately metering electricity use on specific circuits of individual dwellings is one approach to obtaining this information. However, this is extremely costly and hence impractical over millions of individual dwellings. Meters that plug into standard wall outlets are another approach, though these cannot be used with hard-wired appliances such as most air-conditioners. An alternative is to use an aggregate whole-of-house metering device combined with a device disaggregation technique to determine individual appliance usage. This is known as Non-Intrusive Load Monitoring (NILM) and has been studied by numerous authors. See for example the reviews by (Abubakar, Khalid, Mustafa, Shareef, & Mustapha, 2017), (Zeifman, 2011) and (Zoha, Gluhak, Imran, & Rajasegarar, 2012).

NILM techniques rely on data sampled at a relatively high frequency to enable the disaggregation based on the electricity load signals. For example, according to Zeifman (Zeifman, 2011), NILM approaches may be classified into low sampling frequency methods (1 Hz or less) and high frequency methods (10kHz to 1 Mhz). The high frequency methods in particular require dedicated (expensive) sampling equipment though high accuracy can be obtained. By comparison, the data
from smart meters is typically only available at a sampling frequency of 0.00055Hz (30 minutes), although the devices themselves may take measurements at up to 0.016Hz (1 minute) and temporarily store the data internally.

In Australia, smart meters are becoming increasingly common with mandatory replacement of non-smart meters across Victoria nearing completion and installation in new builds and replacement of faulty meters ongoing in Queensland, New South Wales, South Australia and Tasmania. Smart meters provide a relatively large amount of data compared with what has been available historically. Generally, this is available to the consumer at 30 minute resolution, and they may also choose to allow access to third parties.

Some authors have investigated device disaggregation techniques applied at progressively lower sampling frequencies. For example, (Farinaccio & Zmeureanu, 1999) used whole of house current and volt measurements at 1Hz combined with a once off sub-metering to train the model to perform device disaggregation. In 2014, Iwayemi et. Al (Iwayemi & Zhou, 2014) used data at 1Hz sampling together with a classification algorithm to do load disaggregation from whole of house energy use data. Koutitas et al. (Koutitas & Tassiulas, 2016) used whole of house data from a current clamp meter at 0.14Hz combined with fuzzy logic and pattern recognition techniques to perform load disaggregation. Su (S., et al., 2016) used whole of house energy use data at 0.0166Hz (1 minute) combined with a Support Vector Machine model to determine air-conditioner on/off status. Their approach required a 2 week training period where air-conditioning on/off status was known.

In this study the aim was to further reduce the sampling frequency to obtain information on air-conditioner usage from whole-of-house metering data recorded at 0.00055Hz (30 minute) resolution. It is reasonable to expect that there will be a trade-off in prediction accuracy; the extent of this trade-off was of interest. Given weather data is important for air-conditioner operation, ambient temperature from the nearest weather station to a given dwelling was also used.

The model, which is based on Gaussian Process Classification applied to a simple 2-dimensional parameterisation of the energy use and ambient temperature space, is described in Section 2. Two separate data sets with ground truth air-conditioner use information derived from circuit level metering were available. The model was developed and trained on a portion of the data from one data-set; this is described in Section 3. In Section 4 a method of generalising the model to any data-set is then described along with the results from testing the generalised model on the remaining portion of the first data-set. Section 5 then describes the testing of this model on the second, independent data-set. Finally the results are discussed more generally in Section 6.

2. Individual dwelling model formulation

The available data from which to build a prediction model consisted of historical total household energy consumption values at 30-minute intervals, the value at the current interval, the timestamp of that measurement (i.e. date and time), and the approximate location (i.e. postcode) of the dwelling. Using the timestamp and location allowed the local weather data (temperature, humidity, wind speed and cloudiness) from the nearest Bureau of Meteorology ground measurement station to be linked to the model. Since accurate irradiance data is generally not available in real time in Australia this information was not included.

There are many ways of combining these variables in a prediction model. The approach taken here was to base the model on two secondary predictor variables, labelled $L_T$ and $L_R$. $L_T$ is an absolute measure of the deviation of the ambient temperature from a reference level considered to be comfortable. If $T$ is the ambient temperature over a given half-hour interval, $T_{1h,p}$ is the average ambient temperature for that hour of the day over the past period of days, $p$, and $T_{n,1h}$ is the reference neutral temperature for that hour of the day, then $L_T$ is defined by:
\[ L_T = \frac{(T - \bar{T}_{h,p}) + (T - T_n|_h)}{2} \]

Here a period \( p \) of days equal to 14 was used and the neutral temperature was defined as \( T_n|_h = 20 + 3 \sin((h - 7)/12\pi) \) to account for the characteristic occupant preference for cooler temperatures at night-time. \( L_T \) is notionally positive for conditions where cooling might be required and negative for conditions where heating might be required. The magnitude of \( L_T \) is greater if the average ambient temperature over the past period is further from the current temperature than the neutral temperature. In effect, this acts like an inertia on the measured ‘discomfort’, making it less for days that would otherwise be more uncomfortable than the recent trend, and greater for days that would otherwise be less uncomfortable than the recent trend.

\( L_E \) is a relative measure of the total energy consumption compared with the energy consumption on a mild day where air-conditioner usage is considered unlikely. If \( E \) is the total energy consumption over a given half-hour interval and \( E|_{h,m} \) is the average energy consumption for that hour of the day over all mild days \( m \), then \( L_E \) is defined by:

\[ L_E = \left| \frac{E}{E|_{h,m}} \right| \]

Here mild days were defined as days where \( \max(T) \leq 24 \degree C \) and \( \min(T) \geq 14 \degree C \). These levels were chosen since they corresponded to minimal a/c use overall across all dwellings while still providing a relatively high number of mild days from which to calculate \( E|_{h,m} \).

For each dwelling, the first step was to construct a classification model that could predict the air-conditioner on/off status at any 30-minute interval given the two predictor values \( L_E \) and \( L_T \). To this end, a Gaussian Process Classification (GPC) approach was chosen, because it is flexible in that it does not require a predetermined function to be assumed for the relationship between the predictor variables and the response variable, and because it provides a probabilistic prediction.

Gaussian process regression models have been commonly used in applications involving the prediction of future electricity usage from current and past usage (Maritz, Lubbe, & Lagrange, 2018). Here a classification model was used as opposed to a regression model since the aim was to predict (binary) on/off status. In addition, the model is applied here to only historical and current information rather than for future projections.

Key parameters for the GPC model are the choice of covariance function, the likelihood function and the inference method. Here the squared exponential covariance function with automatic relevance determination was used with a constant mean function. To estimate the probability of air-conditioner use, the cumulative Gaussian (i.e. the error function) was used as the likelihood function. The Laplace approximation method was used to calculate the posterior for the Gaussian process and evaluate the negative marginal log likelihood of the model (i.e. the goodness of fit). Further details on the GPC method can be found in (Rasmussen & Williams, 2006).

### 3. Model training & initial testing

The model was developed using data from the Residential Building Energy Efficiency (RBEE) study (Ambrose, James, Law, Osman, & White, 2013). Approximately two years of 30-minute interval data was available for 143 dwellings. Most importantly, this data includes both the total household energy consumption values and the a/c circuit energy consumption for validation of the models.

Dwellings were located in the greater Brisbane, Melbourne and Adelaide regions. The total household energy consumption excluded any solar generation or onsite usage. The air-conditioner consumption included that from all air conditioners ‘hard wired’ to the main circuit board. It did not include any heating or cooling appliances that might have been connected to wall outlets.
The approach taken to develop the individual dwelling models was as follows:

1. The nearest weather station to the dwelling (based on postcode) was identified and weather data matched to the smart meter data according to the metering data timestamps.

2. Intervals missing temperature or energy data were removed. In addition, energy values greater than the 99.9th percentile were removed and energy values less than zero were set to zero. Air-conditioner energy-use values greater than the total household energy-use value were set to the total household energy-use value.

3. The ground-truth a/c on/off status was determined as follows: for air-conditioner energy use >200Wh in a 30-minute interval, at least one air conditioner was considered to be on. (Note: this value was chosen so that intervals classified as ‘on’ were likely to correspond to a/c use over a significant portion of the interval, and to avoid false positives where intervals with only standby power use would be classified as ‘on’ intervals. If only one air-conditioner were known to be present on the circuit, then use of a lower threshold may be possible.)

4. For each dwelling, the data was divided randomly into two equal-sized sets: a training set (50%) and a testing set (50%). Data to fit the models was randomly selected from the training dataset such that at least 400 ‘on’ and 400 ‘off’ intervals were selected, subject to the constraint that no more than 2000 of either was selected.

5. For each dwelling, the three GPC hyper-parameters were optimised to find the best model.

6. The model was then evaluated on the entire testing set (approximately one year of data).

The result was 143 models, one for each dwelling. Figure 1 shows plots for 6 of these dwellings indicative of the larger set. The numbers above each figure are a dwelling identification number. These plots show ground-truth air conditioner on (red circles) and air conditioner off (blue circles) status at 30-minute intervals as a function of $L_E$ and $L_T$ for the equivalent of approximately one year of testing data. Contour lines show the classification model-predicted probability of the air conditioner being on in 10% probability increments based on the training data. The highlighted turquoise contour is the 50% contour, which can be used to distinguish on/off. (Note that any contour could be chosen depending on the desire to minimise either false positives or false negatives.)

The behaviour for dwelling 3 (represented in the top-left figure) is consistent with occasional, cooling-only use of what is likely to be a medium size air conditioner (as evidenced by the moderate relative increase in total household power consumption when a/c is in use) at times of moderate-to-high likely discomfort (i.e. $L_T > 10$).

For dwelling 5 (top-right figure), air conditioning is used for both heating and cooling (usage values corresponding to both negative and positive values of $L_T$). A relatively clear separation exists between the two regions at approximately $L_T = 5$, with heating usage being more frequent. Dwellings 86, 108, 134 and 137 also show an apparent separation between cooling and heating behaviour, though with different characteristics.

Dwelling 86 has relatively high energy-use values occurring at mild conditions and not associated with the metered a/c circuit. The magnitude of these values is similar to the magnitude when a/c is in use. Two possibilities are: the air conditioner is quite small and has minimal influence on the usage values, or alternatively, another device is in use during mild conditions when the a/c is not in use.

Dwelling 108 is likely to have a large capacity air conditioner, given the very high values of $L_E$ when air conditioning is in use and the clear separation between in-use and not-in-use states. Usage is frequent when $L_T$ deviates from a small zone around 5. On the other hand, the data for dwelling 134 suggests much more infrequent usage of a small-capacity air conditioner, only for relatively extreme conditions. The smaller capacity makes distinguishing ‘on’ states more difficult, but the fact that usage is confined to more extreme values of $L_T$ alleviates this somewhat.
Finally, dwelling 137 displays an interesting behaviour. There is a distinct zone of relatively high values of $L_T$ for cold conditions that is not associated with usage of the metered air-conditioner circuit. Meanwhile, the heating a/c usage for similar conditions is distinct and corresponds to lower values of $L_T$. Possible explanations may be use or increased consumption by another appliance on cold days (e.g. a hot water booster switched on over a specific period of cold weather), or another heating appliance that is not measured by the metered a/c circuit (e.g. an electric bar heater).

Figure 1 Test dataset showing example a/c usage behaviours for 6 different dwellings. Red symbols indicate 30-minute intervals with a/c usage; blue intervals without usage. Symbols represent the test dataset. Contours show estimated usage probability based on GPC model fitted to training dataset. Highlighted contour is the 80% probability of use.
The overall prediction accuracy on the test data for all of the dwellings is summarised in Figure 2. This plot shows the actual total hours of a/c usage versus the model-predicted hours of use. The R-squared value of a linear fit through the data is 0.97 with a Root Mean Square (RMS) error of 9.2 hrs, indicating a good fit of the individual dwelling models to the data. Note that the models were fit to separate training datasets.

![Figure 2 Comparison of estimated and actual total hours of a/c use on the test dataset using the individual a/c use models.](image)

4. **Generalising the model**

The results above show that the GPC method can be used to build a reasonably accurate model for a given dwelling provided a period of known air-conditioner operation is available to train the model. However, in the general case, no information, or much less information on the a/c operation is available. Hence, it is highly desirable to generalise the models so that, in the ideal case, no a/c usage information is required for a given dwelling.

The approach taken here was to cluster the 143 a/c use models into a representative set of models, and then apply a general technique for deciding which cluster model to use for each dwelling. The performance when the cluster for each dwelling was known *apriori* is also reported (Section 4.1.2). The general approach for assigning dwellings to clusters is described in Section 4.2.

4.1.1. **Approach**

The purpose of clustering is to reduce the 143 individual dwelling models into a small subset of representative models that, in theory, can be used to estimate the performance of any set of dwellings. The clustering approach was as follows;

1. Firstly, the $L_E - L_T$ space was divided up into a grid of 100 points. Here a truncated portion of the space was used to focus the clustering on the region with most of the data. The minimum distance from each grid point to the 0.5 on/off probability line was calculated.

2. A principle component decomposition was then performed to reduce the number of parameter dimensions to a more manageable number: in this case, three.

3. A Gaussian Mixture model clustering based on maximum likelihood estimation was used to group the dwelling models based on the three principle components. Here we used a diagonal covariance matrix (since principle components should be independent) but unshared covariance matrices. Bayes information criteria (BIC) was used to choose the optimum number of clusters. Clustering was repeated 50 times for each assumed number of clusters to ensure a smooth trend. Based on this, six clusters were chosen.
4. An a/c use model was created for each of the six clusters to represent all dwellings in the cluster. The approach taken was to fit a new GPC model using randomly sampled data points from the training datasets for all of the dwellings in the given cluster. To reduce the computational time, 2000 data points in total were selected to build the model for a given cluster, again ensuring that no more than 400 were ‘on’ data points. The resultant models are summarised in Figure 3. Here, the contours show the a/c use probability based on a subsample of the training datasets, and the data points show the actual on/off binary air-conditioner use values: in this case, also for the subsample of the training dataset.

Cluster 2 corresponds to dwellings with the metered a/c device used for heating only. This cluster have only two dwellings and is very distinct. Clusters 3, 4 and 5 capture qualitatively similar behaviours but with generally decreasing frequency of a/c use for both cooling and heating. These clusters are also somewhat distinct. Clusters 1 and 6 correspond to similar behaviours with mostly cooling a/c use only, though the behaviours tend to be more varied.

Figure 3 Contours showing prediction probabilities in 10% increments for the six cluster models. Symbols indicate on (red) and off (blue) ground-truth for the training data subset.

4.1.2. Predictions based on cluster models – cluster known

To test the clustering the six cluster models were used to predict the a/c use of all 143 dwellings for the test data sets. The resultant actual and predicted hours of use are compared in Figure 4. Symbols indicate the assigned cluster for each dwelling. This plot can be directly compared with Figure 2, which shows the same prediction using the 143 individual models. The R-squared value of the fitted line is 0.86 with an RMS error of 27.8 hrs indicating that the accuracy has decreased somewhat.

The models also provide the a/c use status at specific 30-minute intervals. Figure 5 compares the actual recorded 30-minute a/c circuit energy consumption for one dwelling with the predicted a/c use probability using the model developed for the cluster. Red and green symbols indicate incorrect and correct predictions respectively (based on the 50% probability threshold). For many a/c usage intervals over summer, the prediction is reasonably accurate. Incorrect ‘on’ predictions are generally more frequent than incorrect ‘off’ predictions, and many of the incorrect ‘on’ predictions occur just before or just after correct ‘on’ predictions.
4.2. General model for assigning dwellings to a cluster

In the general case an additional step is required to assign dwellings to one of the 6 cluster models without the benefit of any actual a/c use information. A number of approaches may be envisaged such as using characteristics of the building, the a/c system or the occupants. The approach that achieves sufficient accuracy using only the most readily available information is obviously preferable. Initially using just the spatial distribution of points in the $L_E - L_T$ space was trialled. This resulted in an overall R2 value on the test data-set of 0.74. This compares to 0.86 as described in the previous section when the best cluster for a given dwelling was known based on the actual a/c use.

In practise, it may be feasible to request limited information from dwelling occupants, for example to ask an occupant at one or more particular instants in time if their air-conditioner is turned on or off, or alternatively whether they use a/c for both cooling and heating, or perhaps the capacity of their air-conditioner. This might be accomplished, for example, by using a phone or web based surveying application such as the CSIRO’s Energise App (Bainbridge, 2018).

Here we considered just one possible question – whether the a/c is on at a particular instant. Assuming that it is practical to request this, the problem then is to decide when to survey the occupants to maximise the usefulness of the information. That is, the sampling point in the $L_E - L_T$ space that maximises the chances of identifying the best a/c use model. This is a classic problem in information theory. Thus the approach taken was based on minimising the entropy (information) level for the cluster probability distribution $P_c$ as described below.
1. For a given dwelling, an initial uniform probability $P_c = \frac{1}{N}$ was assigned for each cluster model $c$.
2. Next the entropy (information) level was computed as $H_o = -\sum_{c=1}^{N} P_c \log(P_c)$.
3. A potential sampling point $(l_t, l_e)$ for a given dwelling was then chosen and the probability $P_c(l_t, l_e)$ that that sampling point corresponds to a/c use was evaluated for each cluster model.
4. The new cluster probabilities and entropy levels that would be obtained for both true and false values for the sample point were computed:
   \[
P^*_c|\text{true} = P_c \frac{P_c}{\sum P_c}, \quad P^*_c|\text{false} = P_c \frac{(1-P_c) \sum (1-P_c)}{\sum P_c}
   \]
   \[
   H^*_\text{true} = -\sum_{c=1}^{N} P^*_c|\text{true} \log(P^*_c|\text{true})
   \]
5. The new entropy (information) level that would be obtained if a sample was taken at this point was then estimated from:
   \[
   H^* = H^*_\text{true} \sum_{c=1}^{N} \frac{P_c}{N} + H^*_\text{false} \sum_{c=1}^{N} (1-P_c)/N
   \]
   Note that this is an estimate only.
6. An acceptance probability for this sampling point was then calculated according to:
   \[
P_{\text{accept}} = (H_o - H^*)/H_{\text{max}}, \quad \text{where} \quad H_{\text{max}} = -\log\left(\frac{1}{N}\right)
   \]
7. If $P_{\text{accept}} < Rf$, then the sampling point was accepted (i.e. the information would be requested from the occupant). The cluster probabilities were then updated according to the actual a/c status for the sampled point. That is,
   \[
P_c = P^*_c|\text{true} \quad \text{if} \quad p(l_t, l_e) = \text{true}
   \]
   \[
P_c = P^*_c|\text{false} \quad \text{if} \quad p(l_t, l_e) = \text{false}
   \]
8. Steps 2 to 8 were repeated until either the desired number of sample points was obtained, or alternatively the desired probability level was reached.

4.2.1. Predictions based on cluster models – cluster unknown

The above sampling procedure to assign dwelling to clusters was tested on the RBEE test data-set. The number of sample points was varied and the resulting maximum probability that the dwelling was in a particular cluster was recorded along with the overall RMS error for the prediction accuracy of the resulting set of models across all dwellings. These results are shown in Figure 6 below. When the information from 3 sampling points was used to choose the cluster model for a given dwelling, the overall RMS error was the same as when the ‘best’ cluster model was known apriori. In fact, the RMS decreases slightly below this as the number of sampling points increases beyond 3. This is an indication that the method described in Section 4.1.1 to create the cluster models could be improved.

5. Evaluating the model on a different data-set – Yarrabilba data

Although the RBEE data-set was divided into separate test and training data-sets for model development and testing, the two sets were still derived from data from the same set of dwellings. The best way to evaluate the model is to use an entirely different data-set for testing.

The Yarrabilba data-set (Braslavsky & Matthews, 2017) is similar to the RBEE data-set in that it consists of separate electricity circuit measurements over several years for a relatively large number of detached residential dwellings. In this case, data was available over the period 1 October 2016 to 30 September 2017 from 67 dwellings in a new housing estate located south-west of Brisbane. Electricity measurements recorded at 1 minute intervals were aggregated to 30 minute intervals to be consistent with the methods used above. Whole of house consumption was determined by summing together sub-circuit measurements; this effectively removed the effects of solar generation and battery storage. Ground truth a/c use status was determined by comparing the sum of the energy use over a 30 minute interval for all the identified a/c circuits to the threshold level as used previously.
Once again, ambient temperature data from the nearest Bureau of Meteorology reference station was linked to the electricity data.

The same 6 cluster models developed from the RBEE data-set were used to model the a/c use behaviour of all 69 Yarrabilba dwellings. The entropy minimisation approach was used to determine which cluster model to use for a given dwelling. The number of individual sample points where the actual a/c on/off status was known (for example based on asking the occupants) was varied and the influence on the overall RMS prediction accuracy compared. Results are shown in Figure 6.

The RMS error for the Yarrabilba data-set is significantly higher than for the RBEE data and stabilises to approximately 75 hrs. However, similar to the RBEE analysis, data from around 3 sample points provides a significant improvement in prediction accuracy with subsequent additional data points providing little further benefit. For both data-sets the computed probability corresponding to the assigned cluster averaged across all dwellings is also shown. This is in effect, the estimated average probability that the chosen cluster model is the best model out of the available set of models (not necessarily that the model is a good model). This value is almost identical for both cases and consistently increased as the number of sample points increased.

![Figure 6 Comparison of overall prediction RMS error for RBEE and Yarrabilba data-sets as a function of number of known a/c use on/off sampling points.](image)

6. Discussion

The approach taken here was to build a model that provided predictions at the highest frequency allowed by the data. That is, at 30 minute intervals. The resolution that would be useful is likely to vary according to the intended use of the model. For example, providing householders with seasonal statistics on a/c usage may be highly informative and if this was the only aim then perhaps a simpler approach could have been taken. Alternatively, providing information on use over different time of use tariff periods (i.e., peak, off-peak and shoulder) may also be valuable and would require a model with predictions at sub-daily, and ideally, hourly resolution such as developed here. Equally it is possible to envisage network operators and regulatory bodies making use of both the real-time predictions and the long term aggregate statistics. Given this, in this initial investigation we have cast the net broad in attempting to build a detailed prediction model; while the comparison of model performance has been based around the accuracy of the annual use statistics in the first instance.

The method used to construct the two predictor variables $L_T$ and $L_E$ was based largely on intuition and expert judgement. Many other predictor variable constructions are possible, some of which may improve the prediction performance. Even within this formulation, other approaches such as modifying the ‘neutral temperature’ profile or the averaging period over which the moving average ambient temperature was calculated, or even replacing this temperature with an ‘apparent temperature’ measure that includes humidity, could also yield improvements. For example for a
number of the dwelling models the separation between cooling and heating use tended to be centred on small, positive values of $L_T$, suggesting that the mean neutral temperature may be too low. Nevertheless, an important advantage of the two-parameter parameterisation approach over ‘black-box’ techniques is that simple physical meaning is retained within the variables. In applications where it is not just the raw numerical prediction that is of interest but an understanding of how the real variables contribute and interact, and a degree of confidence around the conditions under which the predictions are plausible, this approach has merit.

The final predictions described here for each dwelling in the RBEE and Yarrabilba datasets were based on one of 6 representative a/c use models developed using part of the RBEE dataset. There would be marginal real ‘cost’, be it computational or otherwise, to increase the number of representative models used. That is, it would be straightforward to construct more representative models, possibly from both data-sets, and these models could provide more accurate predictions — though more effort may be required to choose the best model to use for a given dwelling. Alternative methods of clustering dwellings in the $L_E\cdot L_T$ space may be considered, or even the possibility of changing the question that is asked of the occupants to one that provides greater power to identify the most representative model. Modifying the cluster probability estimate so that it also provides information on the suitability of the chosen model given the observations, would also be of interest.

An additional consideration is that the RBEE and Yarrabilba data-sets are, in general, unlikely to be representative of the broader Australia population (consisting of only dwellings built post 2000 in the case of RBEE and even more recently in the case of Yarrabilba). Thus it is suggested that data from a more diverse range of dwellings be included in future work.

The models developed here predict air-conditioner on/off status as opposed to power consumption. This is because the usage behaviour was of particular interest. The simplest method of converting operating time statistics to energy consumption would be to multiple the former by an estimate of the average power consumption of the air-conditioner. The alternative would be to use a regression based approach rather than the classification method employed here.

Finally an important issue relates to the presence of other electricity loads that have a strong dependence with $L_T$ (or ambient temperature) and, critically, that also have high energy consumption relative to a/c. The a/c ‘ground-truth’ for both data-sets was based on the metered circuit(s) identified as being dedicated to an air-conditioner. In many homes, heating and cooling devices such as plug in heaters and portable a/c’s may be connected to General Power Outlets (GPO’s) and use significant amounts of energy. In addition, non-a/c loads for example hot water systems and, to a lesser extent, fridge/freezers, may also have a significant temperature dependence. In fact other appliances can have seasonal variation in use as well, for example if there is a consistent shift in occupant behaviours between seasons (though this is only important if they are also high consumption devices).

In the case of heating/cooling devices connected to GPO’s, it is probably desirable to have consumption from these devices included in the overall a/c use estimate. Since the models attempt to match the ground truth data they strictly do not include this, though because each model effectively averages over many data points, generally from multiple dwellings, it is likely that a high power use value in winter from, say, a plug in wall heater, will be (correctly) assigned to space heating use (although if compared to the ground truth data this will appear as an incorrect prediction). Alternatively, for a given dwelling it is quite possible for high power consumption values in winter that are actually associated with hot water production to be assigned to space heating. In this respect, using one or two other pieces of information when choosing the most appropriate representative model for each dwelling, for example, the type of hot water system in use, and potentially also the dwelling location (which is known) may lead to significant improvements in the prediction.
7. Conclusion
A new approach for predicting air-conditioner usage behaviour at 30 minute intervals from smart meter data has been described. This model can be used to make real-time or historical predictions of individual dwelling air-conditioner usage. Use of smart meter data is likely to be necessary for making predictions across hundreds of thousands or millions of dwellings where installation of more expensive metering is simply out of the question. Such a model can also provide useful insight into typical use behaviours. Although the model may be applied now, continued refinement and, in particular, inclusion of data from a more diverse set of households are suggested to improve the performance and generality.

References


**Acknowledgements**

The work was funded by the Department of Industry and Science through the EUDM project.