

Air-conditioner use prediction via Gaussian process classification

Research report for Work-Stream 4.4

8 June 2018

Goldsworthy, M.



Copyright and disclaimer

© 2018 CSIRO To the extent permitted by law, all rights are reserved and no part of this publication covered by copyright may be reproduced or copied in any form or by any means except with the written permission of CSIRO.

Important disclaimer

CSIRO advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Contents

Contents	3
1 Introduction	4
2 Model development and results	5
2.1 Development of individual dwelling models.....	5
2.1.1 Formulation	5
2.1.2 Model training and testing	6
2.2 Clustering to create representative models.....	9
2.2.1 Approach	9
2.2.2 Predictions based on cluster models	11
2.3 Model for assigning dwellings to clusters	12
2.4 Alternative approach: clustering prior to model fitting.....	14
2.4.1 Predictions based on cluster models	15
3 Predictions for other datasets: CSIRO Victorian pilot dataset	17
4 Summary and next steps	19
5 References	20

1 Introduction

Energy use by air conditioning (heating and cooling) equipment is often a significant fraction of total annual building energy use. Hence, it is also a significant contributor to energy operating costs and emissions. Before making changes to equipment, products, building design or regulations to address or take advantage of this, it is important to understand in more detail how air conditioners are actually used, so that these changes can be targeted to the right consumers to achieve the desired outcomes.

Unless sub-circuit-metering is installed, it is difficult to determine the air-conditioner usage of individual households. Other options – such as plug-in meters, manually recording usage, or high-frequency, whole-of-house metering coupled to device disaggregation methods – are expensive, unsuitable or impractical. Interval meters are becoming increasingly common and provide a relatively large amount of data compared with the data that has been available historically. Interval data is also likely to be available to the consumer, who may also choose to allow access to third parties.

Three examples outlining the usefulness of access to interval data are outlined below.

- Australia's National Construction Code bases its building energy rating on certain assumptions of air-conditioner usage patterns. These patterns directly affect the design of new buildings. Hence, it is important that they are informed by the most accurate data. The model developed here provides better information on how air conditioners are used in residential dwellings, as well as a technique for updating the models if new validation data is available.
- The ability to predict air-conditioner usage – and in particular, peak demand events – is important for managing the electricity network. The model developed here provides an alternative approach to estimating aggregate energy usage when coupled to air-conditioner stock models and weather predictions. The output is therefore a natural complement to the aggregate air-conditioning estimation work of Mazdeh (1), with outputs from both pieces of work eventually enabling improved validation and data gap-filling.
- Most consumers only become aware of their energy use when they receive their quarterly bill. This generally provides no information on which appliances are consuming the most energy or are the most expensive to operate. Having this information would allow consumers to make informed decisions about the appliances they buy or how often they use them. The model developed here is the first step towards providing this information.

The aims of this work are to:

- build a model for predicting *individual dwelling* air-conditioner operation (on/off status at 30-minute intervals) using only whole-of-house, 30-minute interval meter data and the dwelling location
- assess the accuracy of the model using a test dataset that includes ground-truth information
- use the model to predict air-conditioner usage for at least one other dataset and compare the resultant usage statistics with overall air-conditioner use statistics.

Our model has two main limitations. First, it predicts air-conditioner on/off status, not power or energy consumption. Hence, it can only estimate the hours of operation and when this operation occurs. The model is therefore independent of appliance efficiency, and can be coupled with different appliance stock models to predict energy use. However, by itself, the current model cannot provide an estimate of total air-conditioner energy use and cost. Simple methods, such as asking consumers the capacity of their appliance or calibrating the model at a time of known usage, may be used to resolve this. The second limitation is that the model is developed using data from the Residential Building Energy Efficiency study for 143 dwellings. Hence, its generality and representativeness is tied to that of the underlying study.

2 Model development and results

2.1 Development of individual dwelling models

2.1.1 Formulation

The available data from which to build a prediction model consists 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, local weather data (temperature, humidity, wind speed and cloudiness) from the nearest Bureau of Meteorology ground measurement station was linked to the model. Irradiance data is generally not available in real time, but can be obtained several months later from satellite-derived estimates.

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_E .

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, $\bar{T}|_{h,p}$ is the average ambient temperature for that hour of the day over the past period of days, p , and $T_n|_h$ 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 $T_n|_h = 20 + 3 \sin((h - 7)/12\pi)$. 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. This is summarised in Figure 1 for two of the six possible cases.

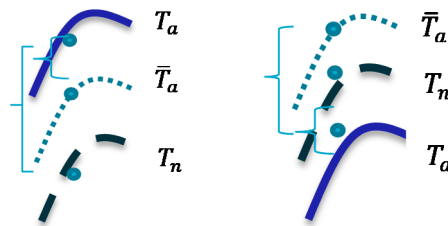


Figure 1 Left: Example with temperature hotter than average over past fortnight, but past fortnight average hotter than comfort neutral; the effective comfort ΔT decreases. Right: Example with cooler than average temperature over past fortnight and cooler than comfort neutral; the average is more than comfort neutral, and the effective comfort ΔT increases.

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 $\bar{E}|_{h,m}$ is the average energy consumption for that hour of the day over all mild days m , then L_E is defined by:

$$L_E = \left| \frac{E}{\bar{E}|_{h,m}} \right|$$

Here mild days are defined as days where $\max(T) \leq 26^\circ\text{C}$ and $\min(T) \geq 14^\circ\text{C}$.

For each dwelling, we seek to construct a classification model that can predict the air-conditioner on/off status at any 30-minute interval given the two predictor values L_E and L_T . 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. It also provides a probabilistic prediction. In a GPC model, constraints are placed on the interaction between the predictor variables (i.e. the covariance function) and on the distribution of possible functions that relate the predictors to the response variable. The distribution of possible functions must be Gaussian.

Key parameters are the choice of covariance function, the likelihood function and the inference method. Here we used the squared exponential covariance function with automatic relevance determination combined with a constant mean function. The squared exponential function is appropriate for a wide range of problems where there is no specific pre-determined pattern expected, such as a period trend or linear increase. The covariance function is given by:

$$k(x_i, x_j | \theta) = \sigma_f^2 \exp \left[-\frac{1}{2} \sum_{m=1}^2 \frac{(x_{im} - x_{jm})^2}{\sigma_m^2} \right]$$

To estimate the probability of air-conditioner use, we used the cumulative Gaussian (i.e. the error function) 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).

2.1.2 Model training and testing

Approximately two years of 30-minute interval data was available for 143 dwellings from the Residential Building Energy Efficiency (RBEE) dataset. This data includes both the total household energy consumption values and the air-conditioner sub-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 sub-circuit energy consumption included usage of 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 is summarised below.

1. The nearest weather station to each postcode was identified and weather data matched to the interval 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 air-conditioner on/off status was determined as follows: for air-conditioner energy use >200 Wh in a 30-minute interval, the air conditioner was considered to be on.
4. For each dwelling, the data was divided randomly into two equal-sized sets: a training set (50%) and a testing set (50%). The training sets were used to fit the models and the testing set was used to test the accuracy of the models.
5. 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 interval was selected.
6. For each model, the three hyper-parameters $(\sigma_1, \sigma_2, \sigma_f)$ were optimised to find the best model.

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

The resultant models are summarised in Figure 2 for 6 representative dwellings. Numbers above each figure indicate the 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 usage of what is likely to be a medium size air conditioner (as evidenced by the moderate relative increase in total household power consumption when air conditioning is in use) at times of moderate-to-high 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 air-conditioning sub-circuit. The magnitude of these values is similar to the magnitude when air conditioning 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 air conditioner 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_E for cold conditions that is not associated with usage of the metered air-conditioner sub-circuit. Meanwhile, the heating air-conditioning usage for similar conditions is distinct and corresponds to lower values of L_E . 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 air-conditioner sub-circuit (e.g. an electric bar heater).

The overall prediction accuracy on the test data for all of the dwellings is summarised in Figure 3. This plot shows the actual total hours of air-conditioner usage versus the model-predicted hours of use. The R-squared value of a linear fit through the data is 0.97, indicating a good fit of the individual dwelling models to the data. Note that the models were fit to separate training datasets.

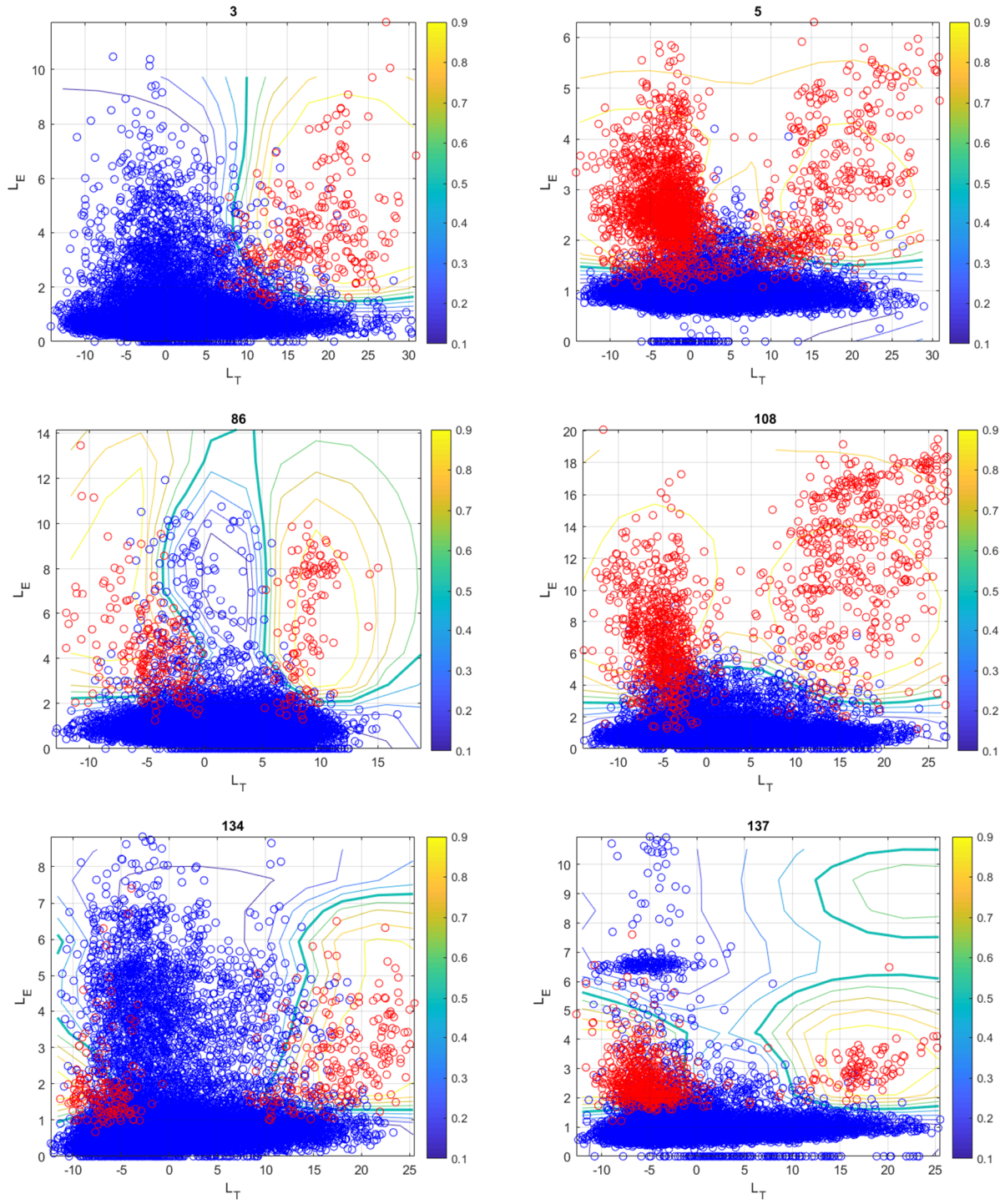


Figure 2 Test dataset showing example air-conditioner usage behaviours for four different dwellings. Red symbols indicate 30-minute intervals with air-conditioner usage; blue intervals without usage. Symbols represent the test dataset. Top left: dwelling with occasional usage for more extreme conditions. Top right: dwelling with frequent use, almost independent of ambient conditions. Bottom left: dwelling with rare usage of a large air conditioner only for more extreme conditions. Bottom right: dwelling with consistent usage for moderate to extreme conditions. Contours show estimated usage probability based on Gaussian process classification model fitted to training dataset. Highlighted contour is the 80% probability of use

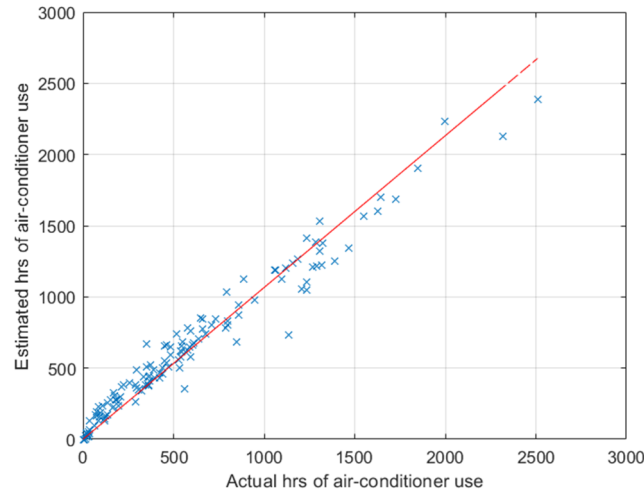


Figure 3 Comparison of estimated and actual total hours of air-conditioner usage on the test dataset for the individual models of dwelling air-conditioner usage

2.2 Clustering to create representative models

2.2.1 Approach

The purpose of clustering is to reduce the 143 individual dwelling models into a small subset of representative models that can be used to estimate the performance of the dwellings. This then allows any dwelling to be modelled, provided that a method is available for deciding to which cluster the dwelling belongs (see Section 2.3). The clustering approach used here and results are described below.

1. The $L_E - L_T$ space was divided up into a grid of points. Here we used a truncated portion of the space $0 \leq L_E \leq 8, 0 \leq L_T \leq 20$ to focus the clustering on the region where most of the data lies. The spacing chosen was 10x10, leading to 100 points. (Note that the method is insensitive to increasing the number of points.)
2. The minimum distance from each grid point to the model contour line corresponding to the desired delineation between on/off was calculated. Here the 0.5 contour line was used.
3. A principle component decomposition was performed to reduce the 100 parameter dimensions to a more manageable number: in this case, three. A plot of the regions contributing most significantly to the three components is shown in Figure 4.

Dimension 1 -> negative L_T (40% of variance)

Dimension 2 -> positive L_T (38% of variance)

Dimension 3 -> high L_E (8% of variance)

4. 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 (sense principle components should be independent) but unshared covariance matrices. Since the number of clusters must be input to the method, this number was varied, and 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.

The variation of BIC with number of clusters is shown in the right plot of Figure 4. The lower the value of BIC, the better (more distinct) the clustering. Based on this, four clusters were chosen. Plots of the three principle components for each dwelling are shown in Figure 5. Symbols indicate

the assigned cluster for each dwelling. Cluster 4 is distinct and corresponds to dwellings with the metered air-conditioning device used for heating only. There are only two dwellings in this cluster. Clusters 2 and 3 are also distinct, with a relatively tight grouping. These correspond to both heating and cooling usage. Cluster 1 is spread out and less distinct and corresponds to mostly only cooling use of the air-conditioning device.

5. A model was created for each of the four clusters to represent all dwellings in the cluster. The approach taken was to fit a new Gaussian classification 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 'on' data points.

The resultant four models are summarised in Figure 6. Here, the contours show the air-conditioning usage 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.

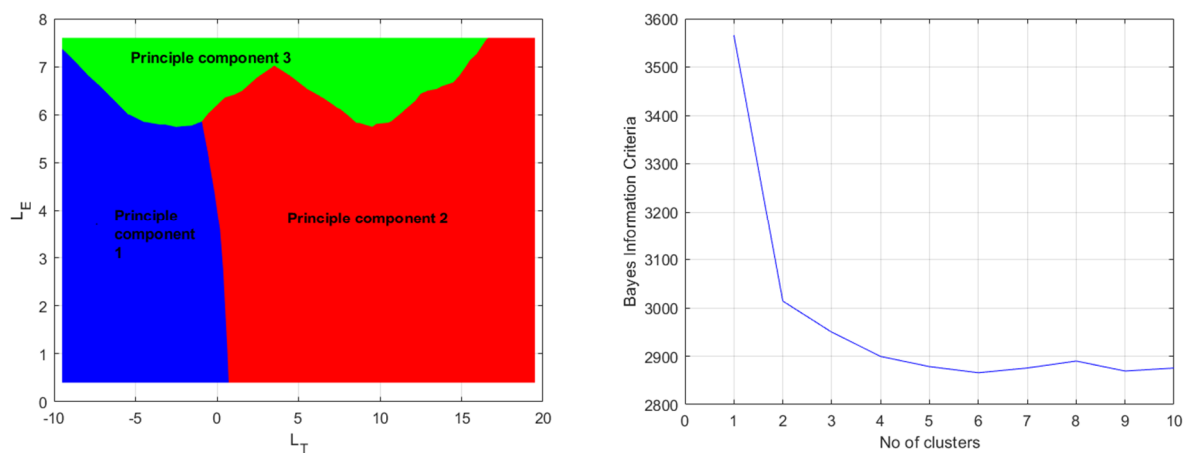


Figure 4 Left: diagram showing regions contributing most significantly to the three principle components used to distinguish different air-conditioner usage model types. Right: plot of Bayes information criteria as a function of the number of different clusters assumed; a lower number indicates a better overall representation of the data

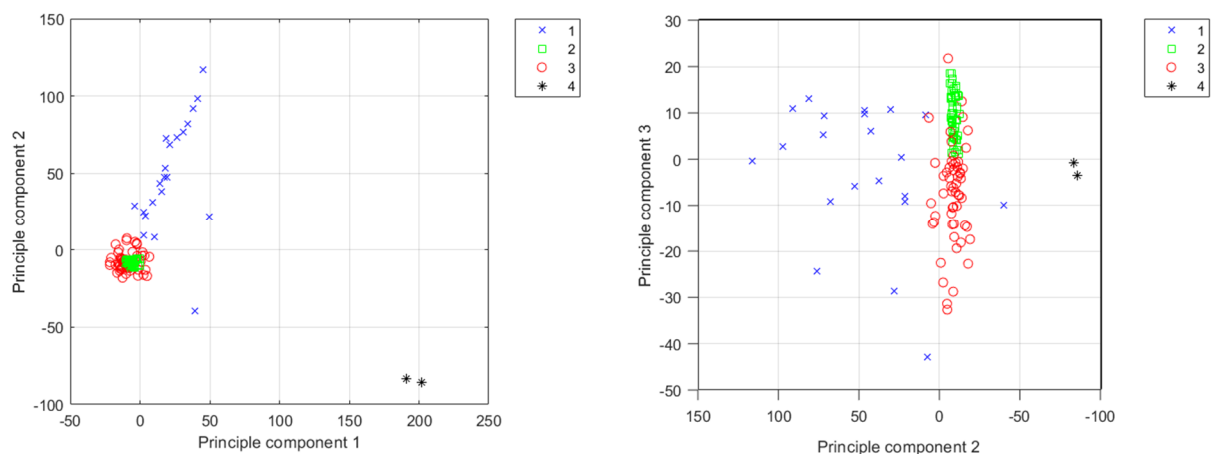


Figure 5 Plots showing grouping of individual dwelling models into four clusters based on the three principle components

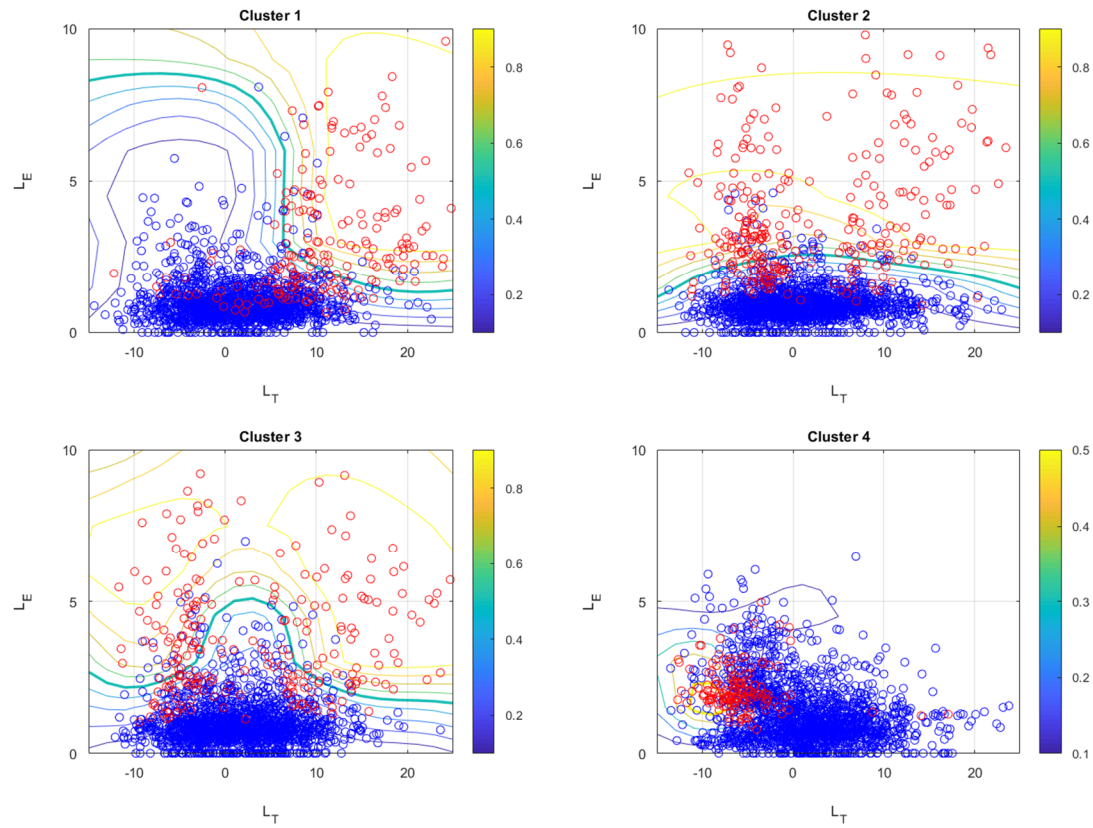


Figure 6 Contours showing air-conditioner use prediction probabilities in 10% increments for the four cluster models. Symbols indicate on (red) and off (blue) ground-truth values for the training data subset

2.2.2 Predictions based on cluster models

In the next step, the above four cluster models were used to predict the air-conditioner use of all 143 dwellings for the test data sets. The resultant actual and predicted hours of air-conditioner use are compared in Figure 7. Symbols indicate the assigned cluster for each dwelling. This plot can be directly compared with Figure 3, which shows the same prediction using the 143 individual models. The R-squared value of the fitted line below is 0.84.

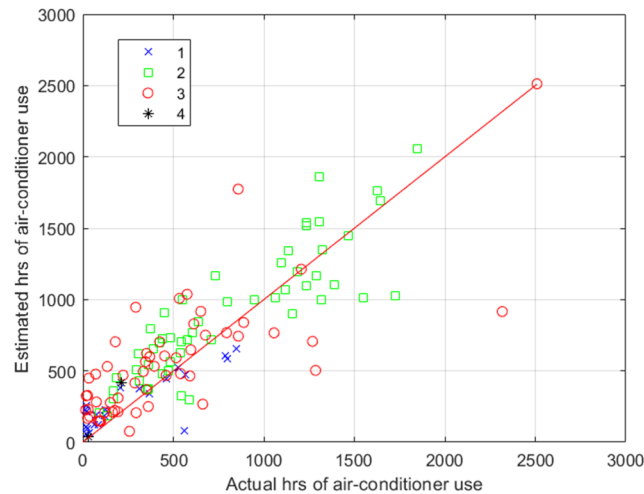


Figure 7 Comparison of estimated and actual total hours of air-conditioner usage on the test dataset using one of the four cluster models to model each dwelling's air-conditioner usage

The models also provide the air-conditioner use status at specific 30-minute intervals. Figure 8 compares the actual recorded 30-minute air-conditioner sub-circuit energy consumption for one dwelling with the predicted air-conditioner use probability using the model developed for the cluster (in this case, Cluster 1). Red and green symbols indicate incorrect and correct predictions, respectively, based on the 50% probability threshold. For many air-conditioner 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.

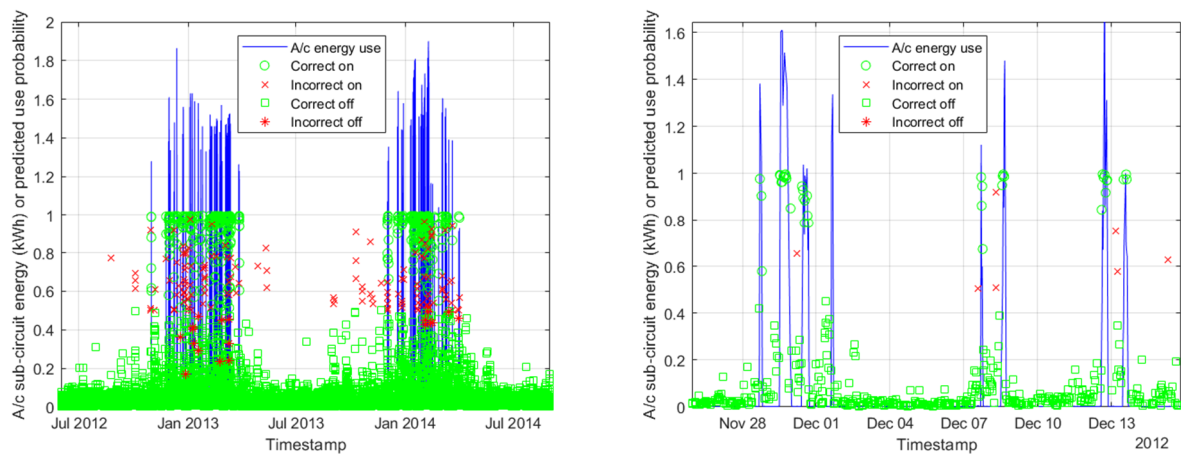


Figure 8 Comparison of air-conditioner sub-circuit energy use (blue lines) for dwelling 19 with prediction made using the model for Cluster 1 using the test dataset. Left: entire date range, right: close up of a selected period

2.3 Model for assigning dwellings to clusters

The final stage of model development involved developing a method for assigning dwellings to one of the four clusters. A large number of variables relating to the dwelling and occupant-reported behaviours were available from the RBEE surveys.

First-order predictors can be considered those that (may) directly relate to the occupant decision to use air conditioning at a particle point in time. For example, the indoor temperature and whether or not occupants are at home. Second-order predictors relate directly to the upper limit of the energy used when in use, for example the size of the conditioned space and the cooling/heating capacity of the air-conditioner. The

magnitude of this upper-limit has an effect on the ability of the models to identify air-conditioner energy usage from the total energy consumption of the dwelling, with higher-consumption devices much more easy to distinguish. Third-order parameters influence the actual energy use at a particular time and may include parameters such as the efficiency/age of the appliance or the effectiveness of the building insulation. Most of the parameters included were second or third-order parameters. These included:

- heating, cooling and hot water system type
- annual electricity and gas spend in ranges
- climate zone
- dwelling theoretical star rating, level of insulation and estimated total heating, ventilation and air conditioning load
- conditioned floor area
- number of occupants (adults and children) and typical occupancy pattern
- occupant-reported typical appliance use frequency in simple categories (dishwasher, washing machine, clothes dryer, cooler, heater), and average winter and summer temperatures in the dwelling
- instances where occupants try to limit their energy use, use fans instead of air conditioning, shut blinds in summer to reduce heat load, choose an air-conditioning set-point that reduces energy use, or purchase green power.

The following parameters derived directly from the energy-use profile were also considered:

- average dwelling energy use
- 25th and 75th percentiles of dwelling energy use
- polynomial coefficients for daily energy-use variation with maximum temperature of the day
- density of points in different regions of the L_E - L_T space.

Unfortunately, none of the above variables could be used to reliably assign dwellings to one of the four clusters and incorrectly assigning a dwelling to a cluster generally resulted in large prediction errors. The most useful variables were the reported frequency of air-conditioner use and the distribution of points in the L_E - L_T space. However, the resultant model was only able to assign dwellings to one of the five (four plus an additional *no air-conditioner use* cluster) with an accuracy of approximately 50%. There are a number of possible reasons for this.

Although it may appear that there are many predictors from which to build a model, many of the survey parameters are not closely associated with air-conditioner energy use, let alone the action of using air conditioning at all. A number of the survey-recorded parameters are only ternary predictors at best, while some, such as usage frequency of non-air-conditioning appliances, or the purchasing of green power, simply were not useful predictors of behaviour for this dataset.

Some variables that may be expected to be useful predictors were not found to be such, or were found to be less useful than expected. The type of heating/cooling system, or whether a heating or cooling system was present and if gas heating was used, were not clear predictors. The direct reported frequency of air-conditioner use was less useful than expected. There are several likely reasons for this, as explained below.

First, the presence of a cooling/heating system is only a necessary condition for its use, not a guaranteed indicator of use. For example, in the case of reverse-cycle systems, occupants may use only the cooling or only the heating function. Secondly, use of other plug-in electric appliances can make an energy-use profile with no sub-circuit metered air-conditioner appear the same as an energy-use profile with a sub-metered air-conditioner heater. Following on, if the survey respondents in both dwellings report frequent heating use, then this measure's usefulness as an overall predictor is reduced. For heating the presence of gas heaters that usually contribute very little to electricity use but can factor into occupant responses around heater usage in

general further complicates the analysis. Finally the presence of other temperature-dependent loads: most notably in winter with water heating, but also in summer due to fridges and freezers is an added complexity. Particularly for dwellings with a smaller capacity air conditioner, these loads may appear to some extent like a heating or cooling device, which can disrupt (i.e. confuse) the model for assigning dwellings to clusters.

In the absence of additional information to distinguish different dwelling behaviours, one option to improve the accuracy of the model for assigning dwellings to clusters is to simply reduce the number of clusters. The models for clusters 2 and 3 are similar. Combining these two clusters reduces the overall R-squared value across all 143 dwellings by only a small amount to 0.8, compared with 0.84 with the model with four clusters. This increases the accuracy of the model for assigning dwellings to clusters to approximately 70%. However, this is still lower than desirable. Hence, in the next section we investigate an alternative approach to model development that clusters dwellings prior to model development.

2.4 Alternative approach: clustering prior to model fitting

The approach described above involved developing individual models of each dwelling, clustering dwellings with similar models, creating a representative model for each cluster, and then building a model that assigns new dwellings to a cluster without any air-conditioner usage information. The advantage of this approach is that it ensures dwellings with similar air-conditioner behaviour are grouped together. In addition, it can also be used to develop very accurate prediction models for individual dwellings, provided some ground-truth data is available to train the model for that dwelling. The disadvantage is that the final step of assigning new dwellings to a particular cluster is not always easy, or requires additional information from the dwelling, as was apparent in Section 2.3.

An alternative approach is to perform clustering up front using information that does *not* include the air-conditioner usage behaviour, and then construct an air-conditioner usage model for each cluster. The advantage of this approach is the process of assigning new dwellings to a particular cluster is simpler. However, it relies on the upfront clustering working correctly to group dwellings with similar air-conditioner usage behaviours.

The approach taken is summarised below.

1. Steps 1-5 in Section 2.1.2 were completed to create the datasets.
2. The L_E - L_T space was divided into six fixed, predefined regions. For each dwelling, the fraction of points in the training data set that fell into each region was determined. Note that no information on whether the points were air-conditioning on or off was used in this step.
3. Gaussian mixture model clustering was used to group dwellings with similar values of the above fractions. Once again, BIC were used to select the optimal number of clusters, which was found to be five. Four of these clusters (2, 3, 4 and 5) were found to be reasonably distinct, as shown in the silhouette plot in Figure 9 (left). (Note that high positive silhouette values indicate that a particular dwelling is uniquely associated with that cluster. Negative values indicate that the dwelling could have been assigned to a different cluster.) The number of dwellings assigned to each of the five clusters is shown in Figure 9 (right).
4. For each cluster, randomly sampled data points from the training data subsets (those created in Step 5 of Section 2.1.2) were combined from all dwellings in that cluster. To reduce computing time, 2000 data points in total were selected for each cluster. The GPC method was then used to build the model for each cluster with optimised values of the three hyper-parameters ($\sigma_1, \sigma_2, \sigma_f$). The resultant five models are summarised in Figure 10. Once again, contours show the air-conditioner usage probability based on the subsample of the training datasets. In these figures, data points have been omitted, because there is a generally a greater overlap of 'on' points in 'off' regions and vice versa, making it more difficult to see the contour lines.

- The cluster models developed using the training datasets were used to predict the air-conditioner on/off behaviour for each dwelling for the test datasets.

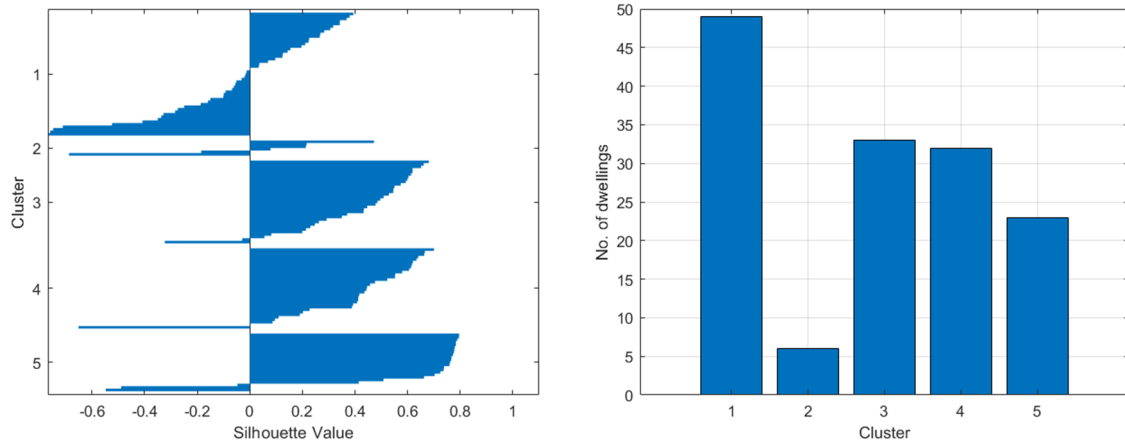


Figure 9 Left; Silhouette plot showing how well the dwellings are grouped into clusters. Positive values indicate dwellings that strongly associated with only one cluster. Negative values indicate dwellings that could have been assigned to a different cluster. Right: Count of the number of dwellings assigned to each cluster

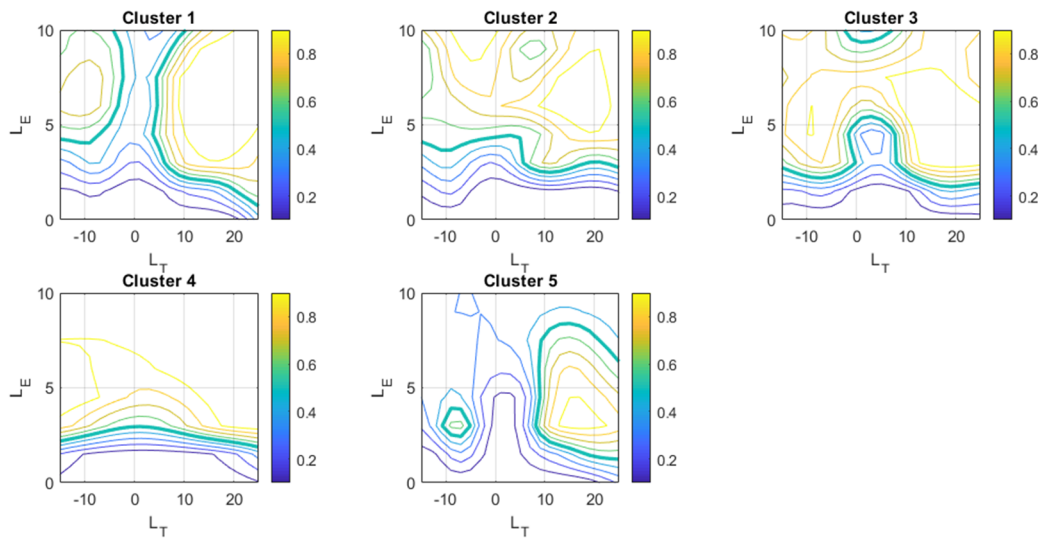


Figure 10 Contours showing air-conditioner use prediction probabilities in 10% increments for the five cluster models constructed using the alternative approach

2.4.1 Predictions based on cluster models

Figure 11 (left) compares the resultant estimated hours of air-conditioner use with the actual values for the 143 households as predicted using the above five cluster models. Different symbols indicate the cluster to which the dwelling belongs. The R-squared value of the straight line fit through the data is 0.68. This is lower than the value obtained in Section 2.3 using the first method (0.84), which indicates a less accurate fit. In general, dwellings with high/low air-conditioner usage are identified as such. The predictions for Cluster 3 appear to have the most scatter, with a small number of dwellings estimated to have much higher numbers of operation hours than in the actual data set. However, as discussed in Section 2.3, one explanation is that these dwellings have significant energy use in cooling and/or heating due to plug-load appliances.

Unfortunately, the data is not available to confirm this. Figure 11 (right) shows box-whisker plots of the estimated number of hours of use for the dwellings in each of the clusters.

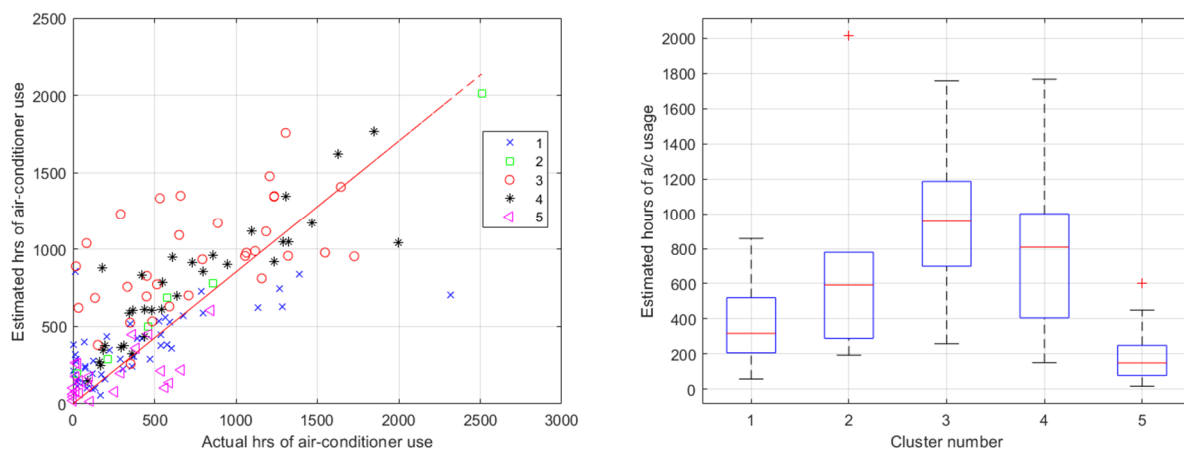


Figure 11 Left: Comparison of estimated and actual total hours of air-conditioner usage on the test dataset using one of the five cluster models to model each dwelling's air-conditioner usage

3 Predictions for other datasets: CSIRO Victorian pilot dataset

Although the analysis described above used separate sets of data for model development and model testing, both were derived from the RBEE study. Validation of the models with a totally different data set is highly desirable. Unfortunately, there are very few interval meter datasets with the necessary air-conditioner use ground-truth data (e.g. a sub-metered air-conditioner circuit) to achieve this.

The CSIRO Victorian pilot survey dataset does not include a separate metered air-conditioner sub-circuit, but it does have respondent-reported number of hours of air-conditioner usage on a specific (known) date. Using this information combined with the overall dwelling consumption interval meter data and linked weather data, we can compare the model-predicted number of hours of air-conditioner usage on the specific date with the respondent-reported hours of usage.

Data from 125 dwellings across Victoria was available. This consisted of interval meter data for one year between October 2016 and November 2017. Respondents reported their hours of air-conditioner usage on the day immediately prior to the date upon which they completed the survey. For most respondents, this was between January and March 2017, though some responses were also obtained in May and July 2017. The alternative model (described in Section 2.4) was used to make air-conditioner usage predictions. Dwellings were first assigned to one of the five clusters based on the Gaussian mixture model clustering using one year of data plotted in the $L_E - L_T$ space. The number of dwellings assigned to each cluster is shown in Figure 12 (left). The corresponding cluster model was then used to predict the air-conditioner usage over a year for that dwelling. A boxplot showing the predicted annual usage for dwellings in each cluster is shown in the right plot of Figure 12.

The predicted usage on the date corresponding to when occupants reported their usage is compared with the respondent-reported usage in Figure 13. Respondents reported usage as one of eight categories; here, the categories corresponding to occupants reporting not being at home and not knowing their usage were combined into the 'no reported usage' category.

The results show there is generally a good agreement between the predicted usage and the respondent-reported usage; note, however, that the categories of reported usage are broad. This gives some level of confidence that the model gives reasonable predictions on an independent data set. For the 6–12 hour category, the prediction is less than the reported, although there are only five or fewer dwellings in each of the four most frequent use categories. Preferably, more data points would be available to perform this analysis.

As a final point, it is expected that the actual hours of usage, as characterised by the sum of 30-minute interval energy being above a threshold, would be equal to or less than the reported hours of usage. This is because once the air conditioner has been switched on for a period of time and the building has reached comfort conditions, it is possible that under mild conditions, the air conditioner may use minimal or no energy for some 30-minute intervals (e.g. if air conditioning is left 'on' overnight and the outside temperature becomes close to the set-point). This is more likely to occur as the length of time in which air-conditioning is reported as 'in use' increases.

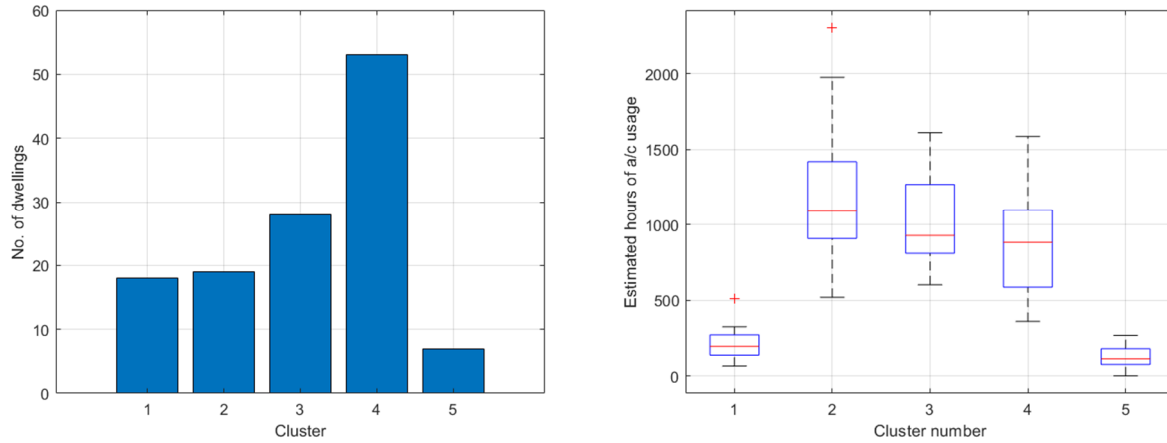


Figure 12 Left: estimated number of dwellings in each air-conditioner usage model cluster for Victoria Pilot dataset. Right: boxplot showing estimated hours of air-conditioner usage by cluster

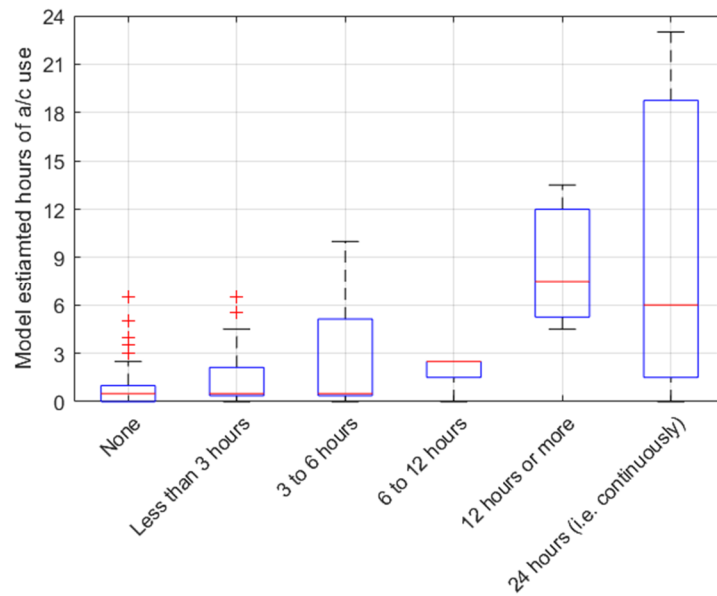


Figure 13 Comparison of model estimated hours of air-conditioner usage with respondent-reported hours of usage on the specific date corresponding to when respondents reported usage

4 Summary and next steps

The air-conditioner usage models developed here may be used in their current form to make real-time or historical predictions of individual dwelling air-conditioner usage at half-hour intervals given interval meter data and the location of the dwelling. However, further refinement of the model could improve the accuracy of the prediction and provide more insight into typical usage behaviours. In particular, the following issues are suggested for consideration.

- The method described in Section 2.4 to cluster dwellings in the L_E - L_T space was relatively simple and used the density of points in pre-defined regions of the space. Alternative methods for clustering dwellings in the L_E - L_T space that achieve a more distinct separation of air-conditioner usage behaviours could be investigated.
- The 'neutral temperature' profile and the averaging period 'p' over which the moving average ambient temperature was calculated were defined based on experience. The dwelling models indicate that the separation between cooling and heating behaviours tends to be centred on small, positive values of L_T , suggesting that the mean neutral temperature may be slightly too low. Alternative neutral temperature profiles and values should be investigated. In addition, the ambient temperature measure could be replaced by the apparent ambient temperature measure, which includes a humidity component that makes it more appropriate to comfort-based calculations.
- The air-conditioner usage ground truth in the RBEE dataset does not include heating/cooling devices that might be connected to GPOs. This is likely to be one of the most significant factors complicating model development. The models developed may also underestimate heating and cooling usage because of this. For training on/off prediction models, the air-conditioner energy-use values are not actually required: only the binary on/off status is required. Hence, it may be feasible to gather new ground-truth data on heating and cooling device on/off status (for example from surveys) from which to further train the models. Alternatively, other sources of this information could be investigated, such as more frequent data on respondent-reported hours of air-conditioning use.

5 References

1. **Mahdavi, N. and Braslavsky, J. and Perfumo, C.** *Inferring temperature dependant loads and gross solar generation from aggregate demand data*. Newcastle : CSIRO, 2016.

CONTACT US

t 1300 363 400
+61 3 9545 2176
e enquiries@csiro.au
w www.csiro.au

YOUR CSIRO

Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills for building prosperity, growth, health and sustainability. It serves governments, industries, business and communities across the nation.

FOR FURTHER INFORMATION

Energy

Mark Goldsworthy
t +61 2 4960 6112
e mark.goldsworthy@csiro.au
w [www.csiro.au/en/Research/EF/Areas/
Electricity-grids-and-systems](http://www.csiro.au/en/Research/EF/Areas/Electricity-grids-and-systems)