

# Panel data analytics for estimating the gross solar output of households

L. O'Neil, A. Berry

01/07/2018

## Contents

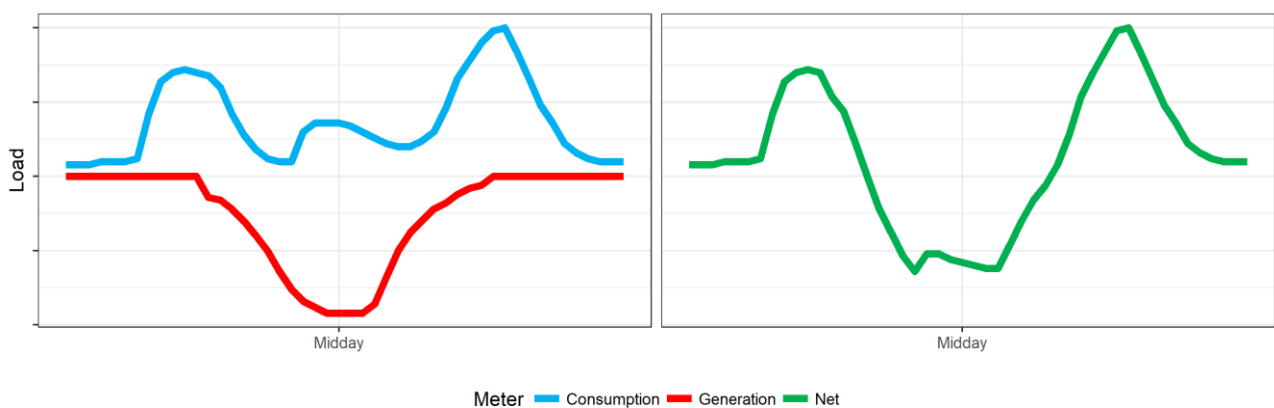
1	The Challenge .....	2
1.1	The problem .....	2
1.2	The need .....	3
1.3	The solution .....	3
2	Progress & Results .....	5
2.1	The training data .....	5
2.2	The model .....	5
2.3	Results .....	6
3	Further Research .....	17
References	.....	18

# 1 The Challenge

## 1.1 The problem

CSIRO has access to half-hourly metered energy data for more than a thousand individual households across many years. Much of this data, however, contains only *net* load measurements, obscuring both the underlying load and generation behaviour of homes with solar PV systems installed. As noted explicitly by AEMO in the development of EUDM work packages this year, such obfuscation of underlying load and generation signals profoundly complicates energy forecasting efforts. Without a clear line of sight into genuine gross PV and consumption, it is profoundly difficult to identify the impact of cloud patterns on instantaneous demand, to track trends in household load behaviour, and to quantify de-rating and degradation effects of deployed solar systems (amongst many other issues).

Solar PV load disaggregation is an attempt to separate the two distinct channels of consumption and generation after they have been combined together into a net load signal. The task is complicated by the fact that net load is driven by a mix of PV system characteristics, instantaneous local weather conditions, building characteristics, appliance uptake and occupant behaviour. By way of example, Figure 1 provides an illustration of the complexity: how one moves from the green curve on the right to the blue and red curves on the left is not at all obvious (the dip around midday could be driven by a reduction in solar generation, an increase in daytime load or some mix thereof).



**Figure 1 - An example of how load characteristics can become hidden when gross load is converted to net load. In this case the consumption hump in the middle of the day is masked, causing the generation to appear flatter and with little indication of a consumption increase**

EUDM presents a multi-pronged approach to this complex problem. In Mazdeh *et al* [1], we explore the identification of regional gross PV output from aggregate load signals. In Zhou *et al*. [2] we look at methods for forecasting individual household PV output based on historical gross PV trends and their relationship to environmental conditions. Here, we focus on bottom-up gross PV estimation based on individual household net load data. Together, these approaches will eventually provide a comprehensive response to gross PV estimation and forecasting. For now, though, the work is preliminary, though promising; providing a roadmap for the path forward.

## 1.2 The need

Though traditional power finger-printing approaches already exist for load disaggregation, these typically require measurements at frequencies of several kilohertz. Current net meters deployed across Australia, however, record at very low frequencies (and are typically reported at sample rates of, at best, every five minutes and, more typically, every thirty minutes). Therefore an approach to disaggregation that is suited to existing smart metering deployments and that does not require additional expensive sensing equipment would be of great benefit. Not only would this save the cost of installing new equipment into many households, but this would provide access to gross signals from historical data, something newly installed equipment could not do and which is critical for understanding real-world solar system degradation. It is therefore CSIRO's goal to present research which can work with existing whole-of-house net smart metering and the coarse temporal data it provides.

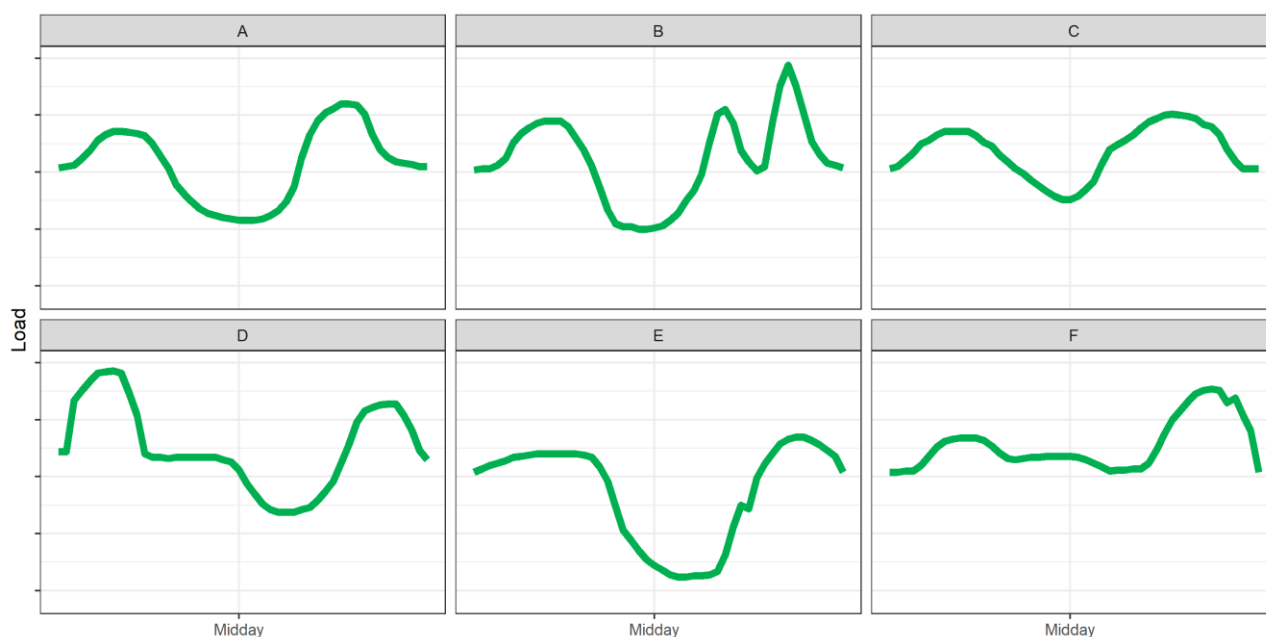
## 1.3 The solution

One natural solution to solar disaggregation at low frequencies is ordinary least squares regression (OLSR). In this approach we could pick several predictors such as time series load features or weather measurements (such as temperature or solar irradiation), and compute the best fit model for the gross solar PV output. However, the nature of the net metered data makes OLSR a non-desirable approach, since it violates the covariance of errors assumption underpinning OLSR. Specifically, the data has both time-dependent characteristics and household-dependent characteristics. OLSR assumes the covariance between errors and predictors is zero. Though this is rarely true in reality, the fundamental concept behind this is that errors in sampling are random, i.e. not related to the predictor. In the case of household load however, with data varying through time, household and location, errors in one of those dimensions will be related to predictors from one of the other dimensions. For example, as the solar PV outputs drops off in winter the variability of output in PV might also reduce. This relates the error in PV output (i.e. household-specific behaviour) directly to the time of year (i.e. time-dependent behaviour). This can cause OLSR to fit less optimal models than other approaches.

One method to navigate the OLSR problems is to consider the data as panel data. Panel data is multi-dimensional time series data where observations for several individuals have been made over the same time period. These are the exact characteristics of the metered energy data. In net metered data, each household may have particular characteristics specific to them, but shared variability over a set time period. By considering daily use, there are many 24 hour windows for each household which vary in unique ways for each house, and yet in similar ways between households.

Figure 2 illustrates some of these similarities and differences. Households A, B, C, D and E all show a dip in the middle of the day where their solar production exceeds their load consumption. However, each of these households behaves completely differently, for example some have distinct early morning peaks like Household D, some have extra overnight peaks like Household B and some have flat consumption like Household E. Despite not having a solar valley in the middle of the day, Household F still shares some similar characteristics to the evening peaks of Households A, D and E. So there are ways in which households can have qualities unique to them

and qualities which are the same, but they all vary over time. This is typical for panel data and panel analysis helps to create models which can best capture this behaviour.



**Figure 2 - Six hypothetical individual loads which demonstrate some similarities and differences between households**

Panel analysis aims to apply a process much like ordinary least squares regression across several panels of time series data. Regression aims to create a model from existing data that can successfully predict data that was not used to train the model. Therefore, when performing any regression based modelling, ground-truth values are required so that statistical models can be formed for the underlying distribution. In our application, to build a model which predicts gross PV output, gross PV output data is required.

Thankfully, using Ausgrid's Solar Home electricity data [3] we have access to gross PV and load data for 300 households. In the future, deeper models can be developed should AEMO provide access to data from thousands of deployed gross meters.

Assuming the location of a home is known (to at least the granularity of postcode), in developing our models we can also draw upon weather measurement data, including approximate solar irradiance and ambient temperature data for the surrounding region.

With data linked, the panel analysis can take net load, solar irradiance and temperature data to train gross generation outputs for individual dwellings (see Figure 3).

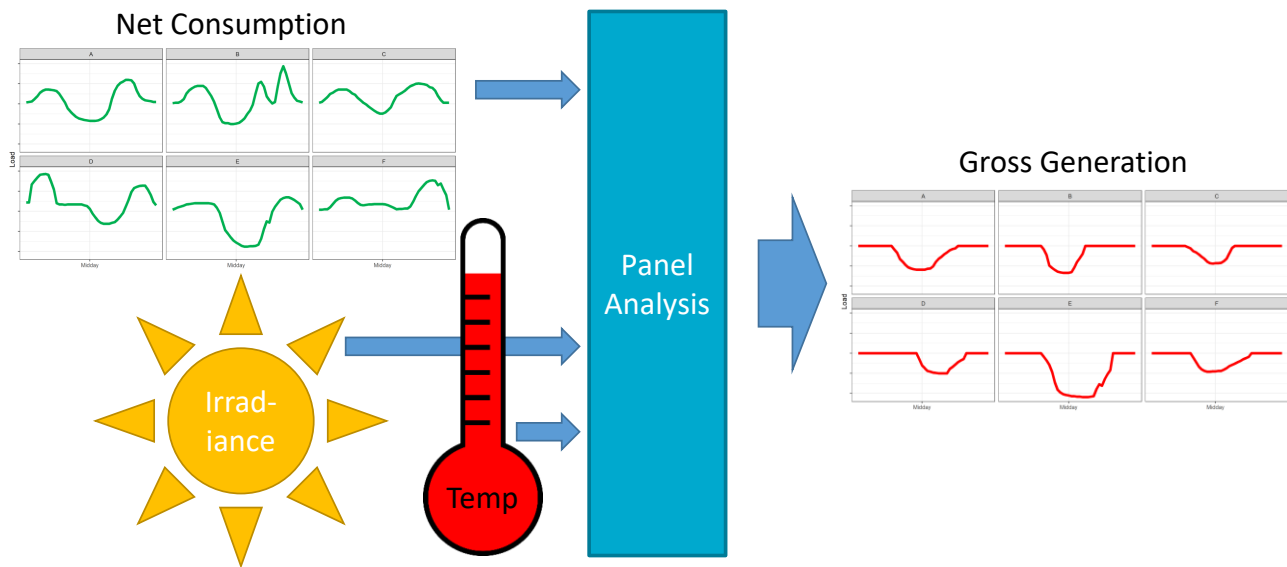


Figure 3 - An illustration of the inputs and outputs to the panel analysis model

## 2 Progress & Results

### 2.1 The training data

As noted in the preceding section, for the development of the panel model, Ausgrid's Solar Home Electricity Data [3] was used. This includes 300 households and their associated postcode. Their data is provided from July 2010 through to June 2013. This modelling focuses on data between July 1<sup>st</sup> 2012 and June 30<sup>th</sup> 2013 for training. Based on household postcode, this data is linked with temperature and rainfall data from the Bureau of Meteorology and solar irradiance data sourced from the Australian Renewable Energy Mapping Infrastructure (AREMI).

### 2.2 The model

An initial panel regression analysis model has been established. This model uses 13 predictors to make half-hourly, per-household, gross generation estimates. The predictors are:

- **Time:** The half-hourly intervals recorded as a decimal in 24 hour time.
- **Net Generation:** Half-hourly net generation readings in kWh.
- **Net Load:** Gross Generation minus Gross Consumption, in kWh.
- **Precipitation:** Half-hourly precipitation in mm.
- **Air Temperature:** Half-hourly air temperature in degrees celsius.
- **School Holidays:** A yes/no flag as to whether a given day is a school holiday for NSW.
- **Type of Day:** A yes/no flag indicating whether a given day is a weekday or on a weekend.
- **Sun Altitude:** The altitude of the sun for the given half-hourly period and postcode.
- **Sun Azimuth:** The azimuth of the sun for the given half-hourly period and postcode.
- **Max Net Generation:** The maximum recorded half-hourly net generation across the year.
- **Max Net Consumption:** The maximum recorded half-hourly net consumption across the year.

- **Solar Irradiance:** The direct normal irradiance (DNI) ( $\text{W/m}^2$ ) for a given half-hourly period and postcode.
- **Overnight Usage:** An estimation of base appliance usage calculated by taking the lowest consumption overnight, each night, between 8pm and 6am. In kWh.

Of these 13 predictors, interaction effects between the following are used in the model:

- **Time** and **Max Net Generation**
- **Time** and **Max Net Consumption**
- **Sun Altitude** and **Sun Azimuth**

The panel regression algorithms then optimise the coefficients for each of these predictors so as to produce an output which resembles the training ground truth data as closely as possible. The interaction effects simply consider those two terms as dependent instead of independent.

Note that the final set of predictors were selected based on multiple training iterations with different training sets. Additional parameters considered during this exploratory phase included day of week (Monday to Sunday) and cloud cover.

## 2.3 Results

### 2.3.1 Results of individual homes

To assess how well the model fits individual homes we will begin by focussing on illustrative case study examples. These examples were chosen as reflecting the performance characteristics seen across the wider dataset, and specifically illustrate the strengths and weaknesses of the modelling approach. For each example, two metrics are presented along with three key visualisations which have been developed. Each aims to look at the goodness of fit of the model from different perspectives. First we will talk through how to interpret these metrics and visualisations while examining House 86.

#### House 86

The Normalised Mean Absolute Error (NMAE) is the mean of all absolute differences between actual values and predicted values in a set, normalised by some number. In our scenario, the NMAE (Annual) is the mean absolute difference of model predicted gross solar PV values and actual gross solar PV values, divided by the maximum actual gross solar PV half-hourly value for that year. The NMAE (Best Days) uses the same normalisation number, but calculates this on only the 12 best days shown in the visual profile of fit in Figure 4. Note that a focus on a subset of the best days is appropriate since a model which can accurately estimate a small number of days could prove useful as an input into other gross PV modelling approaches (such as in the work of Zhou *et al.* [2]). Both metrics are only calculated for daytime hours, that is, 6am to 9pm.

As shown at the top of the summary presentation in Figure 4, for House 86 both NMAE metrics are low, with the best days being almost a third of the annual NMAE. Across a year, the half-hourly outputs are, on average, usually 4.7% off the true value, or approximately 50W.

The top-left vertical graphic in Figure 4 shows the half-hourly differences between estimate and actual gross PV generation for every day of the year. The intent of the visualisation is to highlight

the overall trends in gross generation accuracy over the course of a year. In the figure, the more red represents underestimation and purple overestimation. Note that cell values reflect absolute, rather than normalised differences. For House 86 we can see that very few cells take an extreme value of completely red or completely purple. What we do see is that afternoons are underestimated in the winter months and overestimated in the summer months.

The two stacked horizontal plots on the right hand side of Figure 4 provide a summary of performance across the 12 “Best Days” for each household. A “Best Day” for a particular house is the one with the lowest Least Squares value<sup>1</sup>. The top plot shows the net load curve (with negative values as net consumption, and positive as net generation). We can see that the model produces its best gross PV estimates across a diverse range of net load shapes. The bottom plot shows the actual gross generation curve for each day in black together with the estimated gross generation curve in green. The black curve largely overlaps the green curve for House 86 which is evidence that the model achieves a good fit for the best days, even where gross generation output is intermittent.

Finally the bottom plot in Figure 4 shows a volcano plot of percentiles for House 86. This shows the 50<sup>th</sup>, 65<sup>th</sup>, 75<sup>th</sup>, 85<sup>th</sup> and 95<sup>th</sup> percentiles as well as the maximum values for each half-hourly period. We can view the actual distribution from the real-world Gross PV values on the left hand side, and the predicted Gross PV values on the right hand side. These plots highlight, across a year, that the model captures the overall shape of PV output well. This is critical for determining properties of the PV system such as orientation, rating and localised shading issues (all of which may be used to further tune or refine other gross PV generation estimation models).

We can see Household 86’s shape is biased towards the afternoon, which is captured in both the actual data and the predicted data. This indicates east-facing panels. The magnitudes of the percentiles appear very close as well, further supporting that this model is a good fit for House 86.

In summary, the metrics indicate that House 86 is well predicted by the model, though, there is some seasonal shift in estimates that is not accounted for.

---

<sup>1</sup> Least Squares values are simply a measure of the difference between the actual curve (i.e. the real gross generation data) and the predicted curve (i.e. the model’s estimated gross generation data)

**Max Output:** 0.981 kWh

**NMAE (Annual):** 4.68%

**NMAE (Best Days):** 2.42%

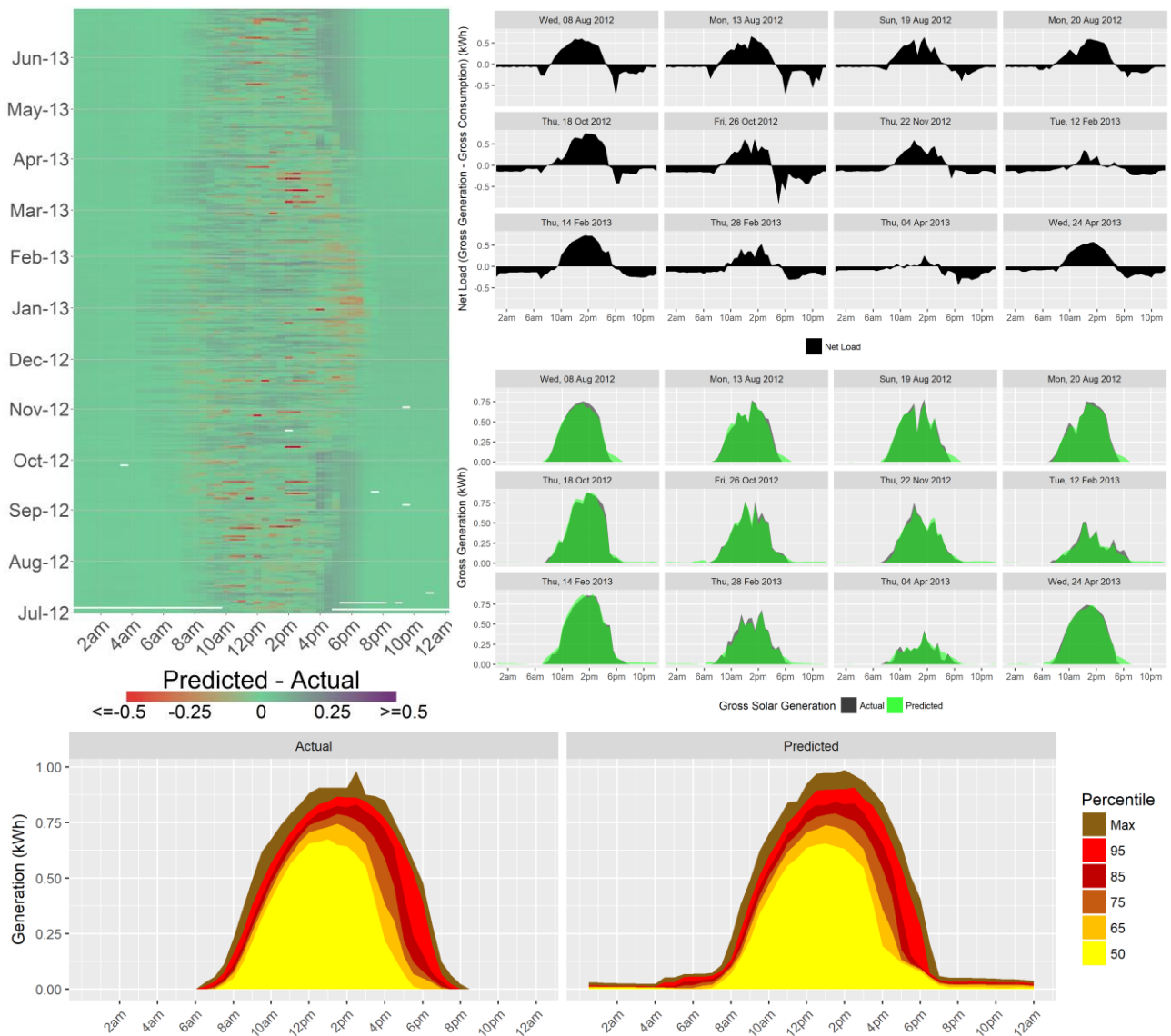


Figure 4 – House 86's profile of fit



## House 107

House 107 (Figure 5) is an illustration of where the model produces lower quality, though still acceptable, gross PV estimates. We can see that though the magnitudes of differences seen in the heat grid plot are typically small, the NMAE performance is about double those seen for House 86. Note that the smaller absolute differences are related to the smaller capacity of the system here when compared to House 86.

We can see from the net load plots that the net generation is much smaller than in our other case examples, however the best days are still a good fit, with the black actual gross generation appearing very similar to the green predicted generation. However, as illustrated in the volcano plots, the typical shape of gross generation is less effectively captured. Specifically, the measured gross generation has an early-morning skew which is not represented in the estimated values. Additionally, the estimated output tends to have lower kurtosis (resulting in peakier output) and overestimates total generation.

In summary, though the model successfully estimates output across a subset of days, it is less reliable across the year as a whole.

**Max Output:** 0.462 kWh

**NMAE (Annual):** 8.04%

**NMAE (Best Days):** 4.25%

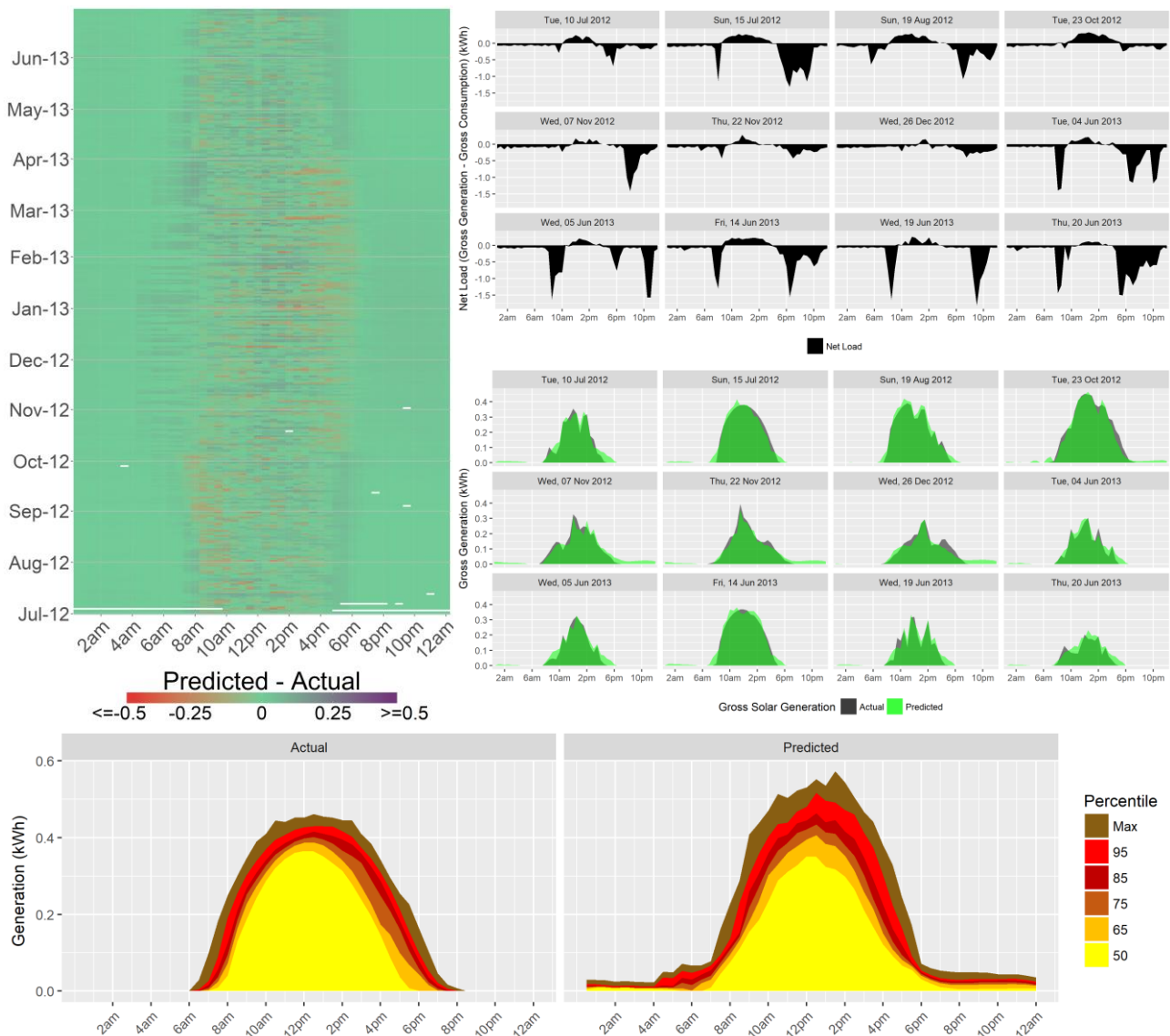


Figure 5 – House 107's profile of fit

## House 127

House 127 (Figure 6) provides an example of where the model output is unsatisfactory. Note that, unlike other houses in the case studies, the home very rarely exports solar PV. Coupled with a small PV system (<1kW) and significant daytime load, the gross PV signal is apparently difficult to identify and likely obscured by variability in load.

The NMAE across the year is high at over 12%. Even on the 12 best days this number is the highest of the case study households (4.6%). The heat grid plot shows that values are regularly underestimated across the year and the volcano plots support this as the percentiles fall well below the actual values.

The volcano plots reveal further issues, as though the morning skew is present in the 6<sup>th</sup> percentile and below, the magnitude is significantly off, and the shape is incorrect in the higher percentiles. At these levels the generation shape is predicted to be much lower in the middle of the day.

Focussing on the best 12 days, a poor fit is also seen across days. The variability of the gross generation is particularly poorly captured. Note, though, that the magnitudes of these estimated shapes roughly outline (or skim) the actual shapes.

It would seem this model fits House 127 poorly, especially when compared to the other households across the case studies. The main point of difference for this household is the high consumption in the middle of the day with very low net generation output, which obscures the gross generation signal.

**Max Output:** 0.613 kWh

**NMAE (Annual):** 12.19%

**NMAE (Best Days):** 4.63%

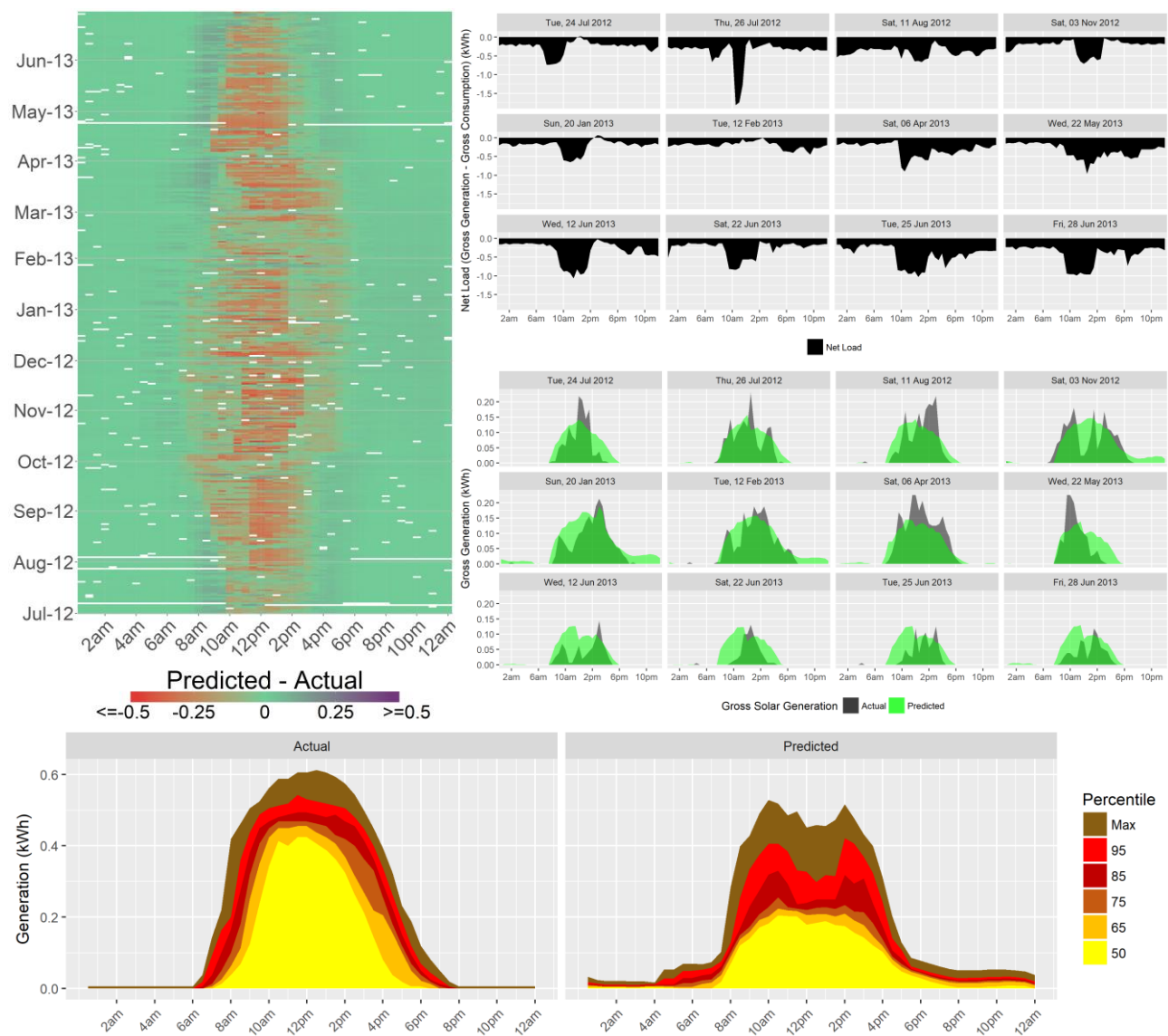


Figure 6 – House 127's profile of fit

## Across the Case Studies

Across the households studied trends begin to appear:

- The estimated gross generation profile on the subset of best days can be highly accurate and, even when the shape is a poor approximation (as in House 127), magnitudes remain close.
- Volcano plots show that the models can capture useful information about output magnitudes and skew in profile shape (even in the poorly performing model for House 127, at lower percentiles the skew in the shape was indicative of the genuine skew). House 107 was an exception as the shape was much too centralised with too low kurtosis.
- Interpretation of reported performance depends upon whether percentage or absolute error is most critical. The difference between NMAEs and heat grids was clear when comparing low output households to high output households, for instance. Low output households are more vulnerable to receiving poor percentage error scores and high output households are more vulnerable to receiving poor absolute error scores.

### 2.3.2 Results of SA3 aggregates

We have now seen that the panel regression model produces both good and bad estimates at the individual household level, but this does not reveal how those estimates interact to affect the quality of regional gross PV estimations. Here, we explore the aggregate gross PV output estimated for 62 homes in Wyong and 27 homes in Newcastle (see Figure 7). This represents a bottom-up approach to aggregate PV estimation that complements the top-down approach of Mazdeh *et al.* (REF).

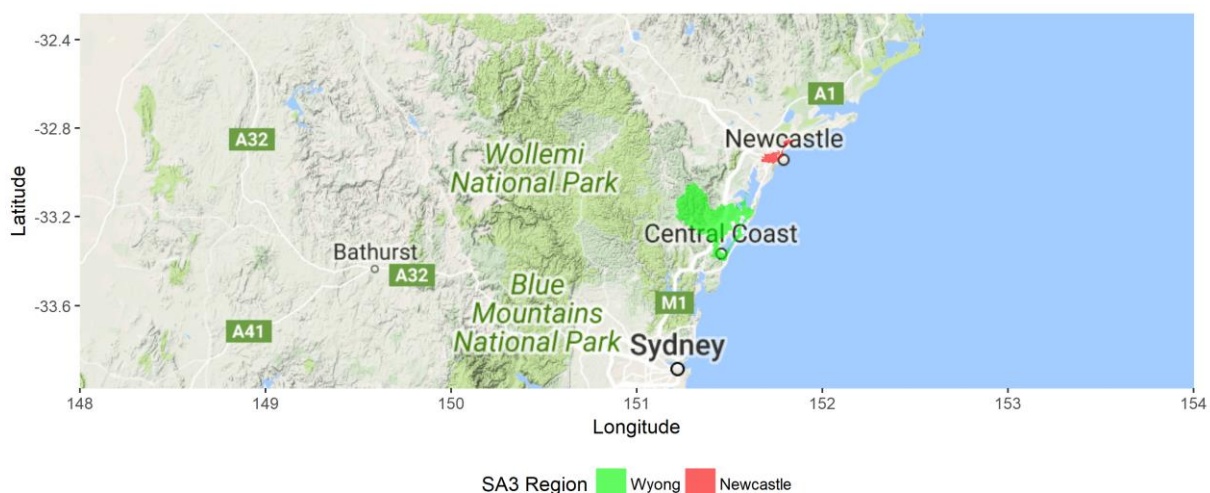


Figure 7 – A map highlighting the two SA3 regions discussed in this report

Figure 8 shows the aggregated heat grids for these two SA3 regions. These cells represent the difference between regional estimated and actual gross solar PV, normalised by the number of houses in that region. If you consider the individual heat grid plots from Section 2.3.1 as having a sample size of 1, they are directly comparable with those in Figure 8. Here we can see the colour of cells are paler than those seen for individual households, indicating that the difference is closer to zero. The seasonal trend of overestimating in winter and underestimating in summer still exists but appears less distinct.

The two calculations of NMAE in Table 1 further emphasise the benefit of aggregation. The Across Region NMAE is an NMAE metric calculated across the region as a whole, calculating the error for each half-hourly measurement of aggregate regional solar PV output and normalising it by the maximum half-hourly PV total output for that region. The Mean Household NMAE calculates the NMAE for each household in the region (as shown in Section 2.3.1) and finds the average of those. Results show that the aggregated estimate has markedly lower percentage error than the individual estimates.

Table 1 - NMAE Summaries for SA3 Regions

SA3	Sample Size	Across Region NMAE	Mean Household NMAE
Wyong	62	4.86%	9.14%
Newcastle	27	1.68%	8.04%

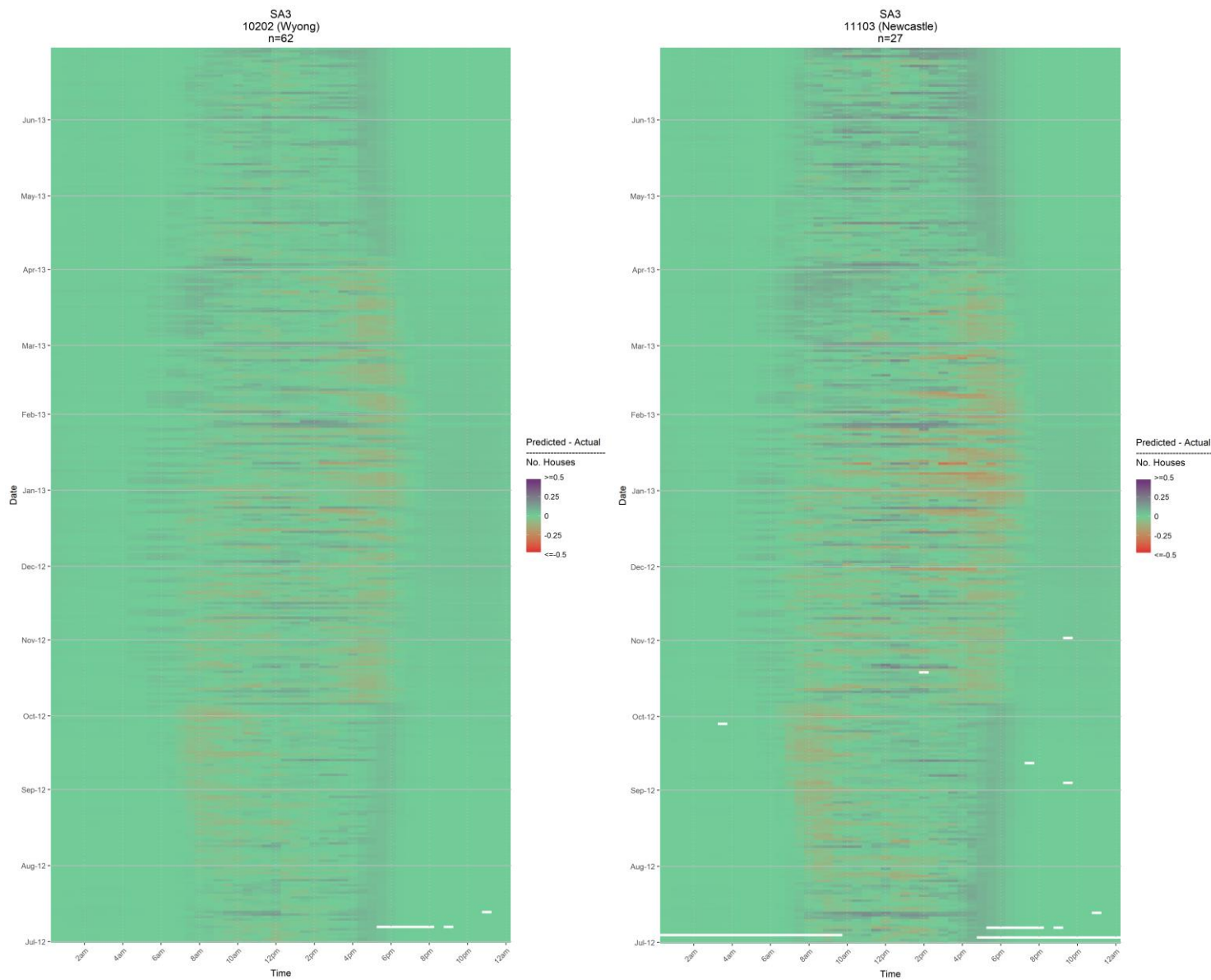


Figure 8 – Heat grid of differences per half-hourly period per house for SA3 regions Wyong and Newcastle. The darker red patches are half-hourly periods where gross generation predictions are under-estimated by a lot and the darker purple patches are half-hourly periods where gross generation predictions are over-estimated by a lot



### 2.3.3 Results of whole set

Previous research conducted into the same area by the University of Melbourne [4] also used the NMAE to evaluate the success of their model. They quoted an NMAE of around 8% across their whole model. Using the definition outlined in Section 2.3.2, our preliminary model without specific tuning to improve under-performing households, has an aggregate NMAE of 4.00% and a mean per-household NMAE of 10.60%. Unfortunately, it is difficult to determine which metric should be directly compared to the University of Melbourne's finding due to a lack of specificity in their reporting. Irrespective of this point, it is clear that the preliminary model developed here is promising and comparable with the state of the art. With further refinement and through integration with models being developed by other EUDM team members, we anticipate further improvement in model accuracy.

When assessing the success of the model through cross-validation (a step important to assure the model will succeed when introduced to new data) results continue to be promising. LS values stay consistently low across multiple tests on unseen data (*folds*), with the majority staying below 0.05 when focussing on the best 12 days only (Figure 9) and below 0.25 when looking at all days (Figure 10). These results suggest that the approach being developed is likely to be well suited to application with new (unseen) data and that the approach does not overfit the model to the training data. It is important to note that, the validation is being applied to homes within NSW however. To ensure that the model is sufficiently robust across climate zones, future training and validation will be expanded to data from other states as well.

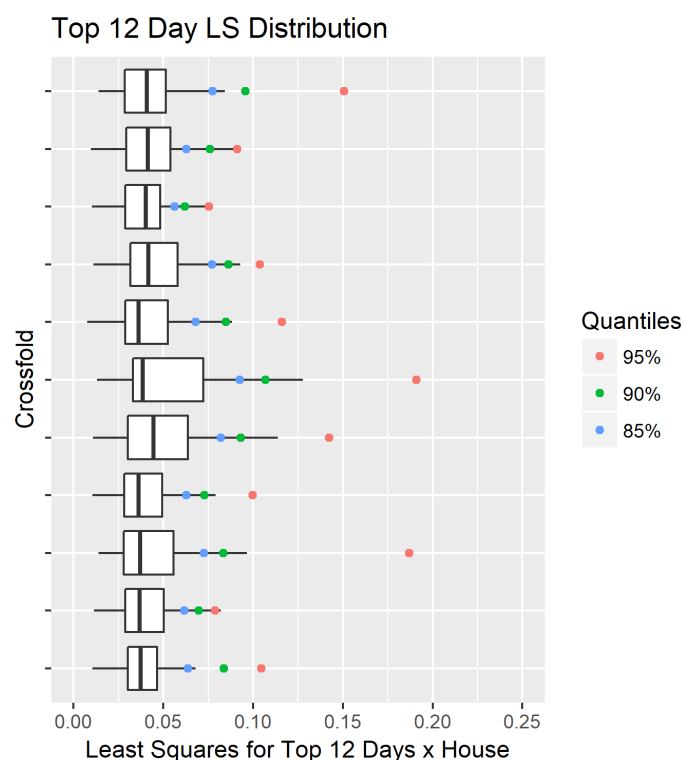


Figure 9 – The distribution of LS for the top 12 days for each household by fold of cross-validation

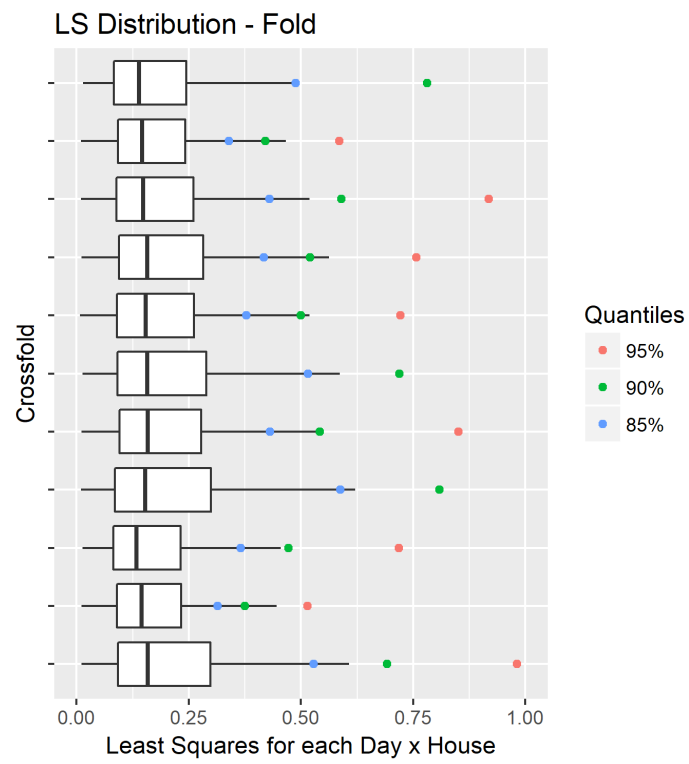


Figure 10 – The distribution of LS for all days for each household by fold of cross-validation



### 3 Further Research

The model is still in its preliminary development stage, but looks promising. For many households, the model effectively capture the shape and magnitude of both daily and annual gross generation outputs. Though there are households where this does not hold, the use of aggregation suggests that individual household prediction errors are effectively smoothed for regional estimation.

Future work will include:

- Where models have underperformed, characterisation of household load and generation characteristics, with a view towards adapting the model to improve performance.
- Expansion of model inputs to include historical time-series data to further improve performance.
- Inclusion of data from other regions to ensure that the modelling approach is performant across climate zones.
- Given that only a subset of days is required to describe key PV system characteristics (such as orientation and sizing), future work will include the development of automated methods to identify the subset of days where the models perform best.

Most critically, the work produced here will be fully integrated into the complementary work being conducted across EUDM to build a robust approach to the estimation and forecasting of gross PV across Australia. As an illustration, the individual household PV estimates produced here can be leveraged to build a full physical PV model for each household through the work of Zhou *et al.* [2], enabling the forecasting of household PV output using only irradiance and temperature data. The characteristics of those models can then inform the Bayesian inference method used by Mazdeh *et al.* [1] to improve their aggregated gross PV estimates. Performance improvements for such aggregations will also come from the fusion of the bottom-up approach developed here and the top-down approach seen in Mazdeh *et al.* [1].

# References

- [1] N. Mazdeh, J. Guo and J. Braslavsky, Disaggregation of zone substation loads across Australia, Newcastle: CSIRO, 2018.
- [2] J. Zhou, Y. Wang, K. Nguyen, T. Hongda and T. Guo, Estimating solar photovoltaic model parameters, CSIRO, 2018.
- [3] Ausgrid, "Solar Homes Electricity Data," [Online]. Available: <https://www.ausgrid.com.au/Common/About-us/Corporate-information/Data-to-share/Solar-home-electricity-data.aspx>.
- [4] V. D. Ruelle, M. Jeppesen and M. Brear, "Rooftop PV Model Technical Report," 2016.
- [5] Australian PV Insitute, "Mapping Australian Photovoltaic installations," [Online]. Available: <http://pv-map.apvi.org.au/historical#4/-26.67/134.12>. [Accessed 14 June 2018].

#### CONTACT US

**t** 1300 363 400  
+61 3 9545 2176  
**e** [csiroenquiries@csiro.au](mailto:csiroenquiries@csiro.au)  
**w** [www.csiro.au](http://www.csiro.au)

#### AT CSIRO, WE DO THE EXTRAORDINARY EVERY DAY

We innovate for tomorrow and help  
improve today – for our customers, all  
Australians and the world.

Our innovations contribute billions of  
dollars to the Australian economy  
every year. As the largest patent holder  
in the nation, our vast wealth of  
intellectual property has led to more  
than 150 spin-off companies.

With more than 5,000 experts and a  
burning desire to get things done, we are  
Australia's catalyst for innovation.

CSIRO. WE IMAGINE. WE COLLABORATE.  
WE INNOVATE.

#### FOR FURTHER INFORMATION

##### **Energy**

Lachlan O'Neil  
**t** +61 2 4960 6103  
**e** [Lachlan.ONeil@csiro.au](mailto:Lachlan.ONeil@csiro.au)

##### **Energy**

Dr. Adam Berry  
**t** +61 2 4960 6123  
**e** [Adam.Berry@csiro.au](mailto:Adam.Berry@csiro.au)