Programme Area: Smart Systems and Heat

Project: WP1 Appliance Disaggregation

Title: Data Quality Report

Abstract:

The ETI collected utility meter and other data (e.g. room temperatures, humidity, and HEMS control data) from five dwellings over a period of six months. Using the collected data, work was conducted to evaluate different machine learning algorithms, research appropriate data features and
calibrations thereof, and test the 'art of the possible'. The work sought not only to understand historical human activity within the building, but also to estimate probabilities of future hot water usage, occupancy and heating needs. This report presents the data quality issues present in the first 10 of 30 hard drives on which data from these homes was stored.

Context:

The High Frequency Appliance Disaggregation Analysis (HFADA) project builds upon work undertaken in the Smart Systems and Heat (SSH) programme delivered by the Energy Systems Catapult for the ETI, to refine intelligence and gain detailed smart home energy data. The project analysed in depth data from five homes that trialed the SSH programme's Home Energy Management System (HEMS) to identify which appliances are present within a building and when they are in operation. The main goal of the HFADA project was to detect human behaviour patterns in order to forecast the home energy needs of people in the future. In particular the project delivered a detailed set of data mining algorithms to help identify patterns of building occupancy and energy use within domestic homes from water, gas and electricity data.

Data Quality Report

CLIENT: ETI

DATE: 09/12/2017

Version History

Version	Date	Description	Prepared by	Approved by
0.1	30/10/2017	Stage 1 draft	Alberto Favaro, Cris Lowery	Oliver Rix
1.0	08/12/2017	Stage 2a final	Alberto Favaro, Zhihan Xu	Cris Lowery

Contact

Name Oliver.Rix@baringa.com +44 7790 017 576

Confidentiality and Limitation Statement

This document is provided to the ETI under, and is subject to the terms of, the Energy Technologies Institute's Agreement for the High Frequency Appliance Disaggregation Project.

Contents

1 Executive Summary

The ETI is collecting utility meter and other data (e.g. room temperatures, humidity, and HEMS control data) from five dwellings over a period of six months. Using the collected data, work will be conducted to evaluate different machine learning algorithms, research appropriate data features and calibrations thereof, and test the "art of the possible". Crucially, the work should not only seek to understand historical human activity within the building, but also to estimate probabilities of future hot water usage, occupancy and heating needs.

Data quality is key to the success of this project and any data quality issues could present a material risk. As such, this report presents the data quality issues present in the first 10 of 30 hard drives, helping identify any risks or methodology changes that will be required in the project. The main risks and complications identified are:

- Ability to generalise has been limited as it will be hard to use 1 property due to the presence of a solar panel;
- There are significant gaps in the data, shortening the useful time period, and requiring us to focus in on smaller time intervals;
- The HEMS data is stored in an unstructured way, leading to extra mapping work and risk of data loss due to syncing;
- There is significant drift in the time series, meaning we will not be able to perfectly line up the data;
- We have only investigated 10 of the 30 hard drives that we will receive and therefore the data quality of 20 hard drives remains unknown.

Overall, data quality introduces some risks to the project and means we have to be selective with the data we chose to analyse as we build up the analysis. However, it is not a blocker, and we recommend proceeding.

2 Introduction

The ETI is investigating the development of a Home Energy Management System (HEMS) capable of optimising the comfort of a dwelling's residents while managing the necessary energy expenditure. As part of this initiative, they are investigating a system that can learn future patterns of occupancy and needs of its residents in terms of hot water usage and heating, using nonintrusive monitoring equipment from two or more utilities. A few key differentiators of the work being currently undertaken, as compared to prior "Non-Intrusive Appliance Load Monitoring" research, are:

1. Monitoring multiple utilities to provide more information and contextual knowledge to better understand appliance usage and location.
2. Potential use of priors to more effectively identify patterns of appliance usage. Statistical properties of appliance usage will feed into these priors, i.e. probability by time of day.
3. Learning workflows and recognising a new work process as part of an existing workflow or as a first work process in one or more known workflows.

To further develop this research, the ETI is collecting utility meter and other data (e.g. room temperatures, humidity, and HEMS control data) from five dwellings over a period of six months. Using the collected data, work will be conducted to evaluate different machine learning algorithms, research appropriate data features and calibrations thereof, and test the "art of the possible". Crucially, the work should not only seek to understand historical human activity within the building, but also to estimate probabilities of future hot water usage, occupancy and heating needs..

Starting from the data collected, Baringa and ASI will jointly support the research activities relating to testing the "art of the possible", which can be broken down into the 5 below (sub)-stages. Throughout the Data Exploration sub-stage, research was conducted on patterns in the data, data quality issues and scripts were developed for processing the data. The output from the stage is predominantly code, but there is also a data output and this data quality report.

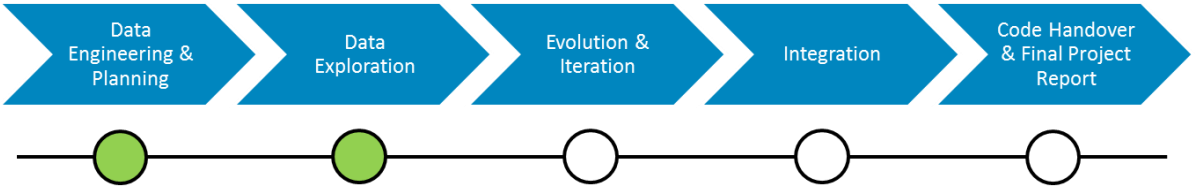


Figure 1: Project stages and sub-stages

3 Overview

The data quality checks described in this report are not intended to be exhaustive, but instead to provide the data scientists with a toolkit to help them identify issues and guide their efforts to the right property and point in time. As such, our understanding of the data quality issues will continue to mature and the report is intended to provide our best current view.

The data quality checks have been designed for each data type individually and are predominantly statistical, due to the significant data volumes, for instance: data completeness, stuck values and load profiles. Using statistical techniques will help pick up on many types of issues and general themes in the data, but will not guarantee that every issue is identified. Additionally, it is not possible to explore every type of potential issue and a pragmatic approach has to be followed, leaving a few small areas of risk. Areas we did not explore include meaningful correlations (i.e. is power correlated to temperature or do calls for heat correlate to gas usage), calibration issues or anomaly detection.

To date, we have received the data specification, the HEMS data schema and 20 hard drives that we believe contain the data for the properties running up to mid-September, and we expect to receive a further 10 hard drives. Our data quality review and analysis has focused on the first 10 hard drives received, which have data from late March to early July, as these are the hard drives that have been uploaded to AWS.

2 of the 10 hard drives' asset numbers did not map back to the asset numbers provided, but having examined the electricity time stamps in the hard drives suggests that it is an asset numbering issue as opposed to a different dataset, which was also confirmed with the ETI. As such, it is assumed that asset number ETI1123 corresponds to property H45 for the time period 31/03/2017-27/04/2017 and that asset number ETI1126 corresponds to property H71 for the time period 28/04/2017 – 02/06/2017. The 10 hard drives also contain two HEMS files, the latter of which we did not audit. The HEMS file we audited had data for 3 properties running up till May. Another key point we found, is that property H20 has a solar panel making it very hard to analyse, and this may be excluded from the work.

	H20	***H25***	***H45***	***H71***	***H73***
No. of hard drives	1	3	3	2	1
Electricity data timespan	31 May – 3 Jul	21 Mar – 20 Jul	31 Mar – 4 Jul	28 Apr – 3 Jul	5 Jun – 3 Jul
Electricity data quality	Solar panels present	Frequent, repeated long gaps	~15 days missing; well-defined gaps	~2 days missing; well-defined gaps	~5 days missing; well-defined gaps
Water	31 May – 3 Jul	21 Mar – 20 Jul (2 water meters)	30 Mar – 4 Jul. Good data quality	28 Apr – 3 Jul. Good data quality	5 Jun – 3 Jul. Good data quality
HEMS database 1	NA	20 Mar – 8 May	20 Mar – 8 May	25 Apr – 8 May	NA
HEMS database 2	Unaudited	Unaudited	Unaudited	Unaudited	Unaudited
Home survey & floor plans	Available	Available	Available	Available	Available

Table 1: Overview of data received by property for the first 10 hard drives

The next four sections represent the key data types and in each section we provide an overview of the data quality checks run and the results thereof. We then continue with a conclusion and recommendations section.

3.1 Summary view

Based on the data completeness view from the initial first 10 hard drives we have prioritised the data we analyse by property and then by date.

1. H45: Has most electricity and water data available, with no major issues
2. H71: Has second most electricity and water data available, with no major issue. Water and electricity consumption appear a bit low
3. H73: Does not appear to have HEMS data (may be present in second database)
4. H25: Deprioritised due to frequent and long gaps in electricity data
5. H20: Deprioritised as it has solar panels, making analysis very difficult.

Our analysis to date has mainly focused on one property, H45, as it was the property with the highest quality as described above. Within that property, we mapped the initial 39 days of data in terms of data completeness and filtered out 15 days due to data quality issues. The below table highlights the better patches of data in green and the 24 days we are currently working on are those highlighted green in the aggregate view. It also bring to life that the two key gaps are due to gas and electricity data.

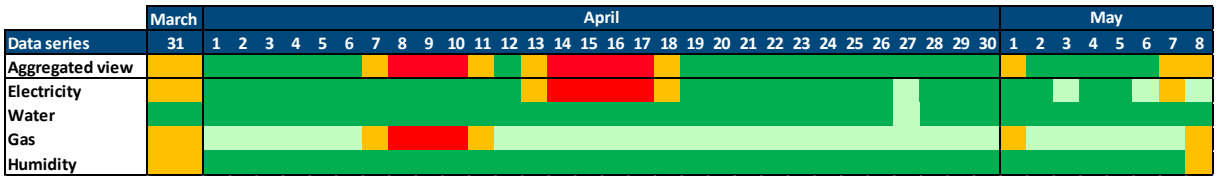


Table 2: Data completeness for H45 for first 39 days of data (green is very high data availability, light green is good data availability, amber is medium data availability and red is no data availability)

4 Electricity Data Quality

4.1 Overview

There are data capture issues leading to gaps in the data as well as data files that are removed due to data quality issues, the former being the larger issue. Multiple thresholds are applied to the electricity data files in order to determine whether a file (c. 30 minutes of data @ 204,918 Hz) is anomalous, and if so, the current view is that we should exclude it from our analysis.

Check	Criterion	Percentage of anomalous files
File duration	The file duration differs from the median by ±3 s	2.8%
Collection duration	The collection duration differs from the median by ±1 s	0.8%
Number of samples	The number of datapoints in the file is 350,000,000.	0.5%
Stuck values	The file is anomalous if it contains any stuck interval longer than 10 'ticks'.	0.1%
Sampling rate	The sampling rate of the raw measurements is at 204,918 Hz.	0.0%
Voltage minimum, maximum and root-mean-square values	The minimum, maximum and root-mean-square values are within 2.5 standard deviations from their means.	0.1%

Table 3: Filter types and percentage of files affected

4.2 Detailed Tests

4.2.1 Logs

The electricity meter generates logs, which we use to filter out files with file durations and collection durations outside of our threshold (1705-1711 and 1707-1709 seconds, respectively).

File duration is the time between a 'file open' and a 'file close' event - recorded when writing the electricity measurements to memory. The distribution of file durations is summarised in the following table:

count	13546.000000
mean	1685.921838
std	222.771588
min	22.407000
25%	1707.969000
50%	1708.187000
75%	1708.297000
max	19911.000000

Table 4: File duration statistics across 10 hard drives

The entries in the table have the following meaning:

- **Count:** the total number of files examined.
- **Mean:** the mean file duration.
- **Std:** the standard deviation of file duration.
- **Min:** the duration of the shortest file.
- **25%:** the 25th percentile file duration (the duration of a file that is longer than 25% of files).
- **50%:** the 50th percentile file duration (the median file duration).
- **75%:** the 75th percentile file duration (the duration of a file that is longer than 75% of files).
- **Max:** the duration of the longest file.

Defining a file as anomalous when the file duration differs from the median by more than 3 seconds in either direction flags 2.8% of all electricity files. While this is not a large number, the file duration was found to be less accurate than one would expect. This is likely to be due to the fact the electricity logs are based on the Network Time Protocol (NTP).

Collection duration is the time between the start and end of electricity measurements. The distribution of collection durations is summarised in the following table:

count	13546.000000
mean	1712.544700
std	181.325563
min	1708.031000
25%	1708.062000
50%	1708.078000
75%	1708.094000
max	19911.156000

Table 5: Collection duration statistics across 10 hard drives

The entries in the table have the same meaning as in the section about file duration. Defining a file as anomalous when the collection duration differs from the median by more than 1 second in either direction flags 0.8% of all electricity files. We observe that collection duration is more consistent than file duration. This is expected, as collection duration is related to the actual time taken for the measurement, whereas file duration depends on the hardware and software used to store the results.

4.2.2 Number of samples

Out of 13,546 files, 26 were removed as the files were empty, and 46 as they were partial files. All complete files have 350,000,000 data rows.

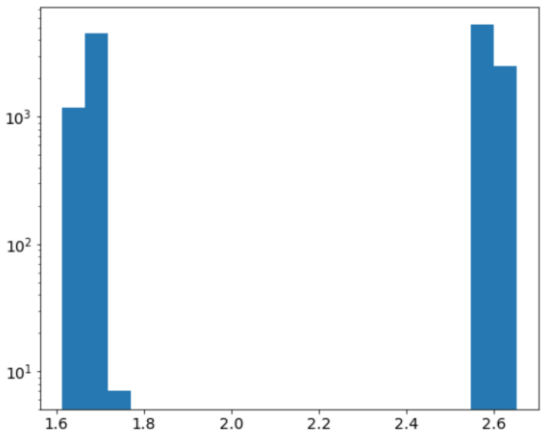
4.2.3 Stuck values

For each file, we looked at stuck intervals of length N = 1, 2... 20, in the voltage channel. In particular, we counted how many sequences of stuck datapoints are present with length N, and then divided this number by the length of the file (350,000,000 datapoints). This returned a percentage of stuck intervals, which is what we display in the histograms below.

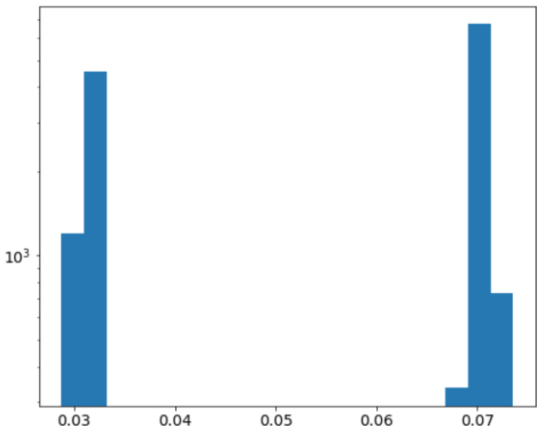
The histograms describe how many files contain a given percentage of stuck intervals. In the first figure, we focus on stuck intervals of length 1, and find that a large number of files contain either about 1.7% or about 2.6% such anomalies. This can be explained by the discretisation that occurs when, in order to store values in memory, the analogue signal from the measurements is converted into a digital signal.

Subsequent figures concentrate on stuck intervals of length 2, 4, and 10. We observe that as the length of stuck interval increases, the number of anomalous files decreases. This is because the discretisation that occurs upon converting from analogue signal to digital signal becomes less of an issue. Eventually, when stuck intervals of length 10 are considered, the quality check always flags the same files. This gives us confidence that the quality of those files is actually lower owing to long stuck values – as opposed to trivial issues of analogue to digital conversion.

Motivated by the above considerations, we have decided on the following criterion for anomaly detection - the file contains a sequence of 10 or more stuck datapoints. This labelled 0.09% of the electricity files as anomalous.



Percentage of stuck intervals having length 1 with respect to file size.



Percentage of stuck intervals having length 2 with respect to file size.

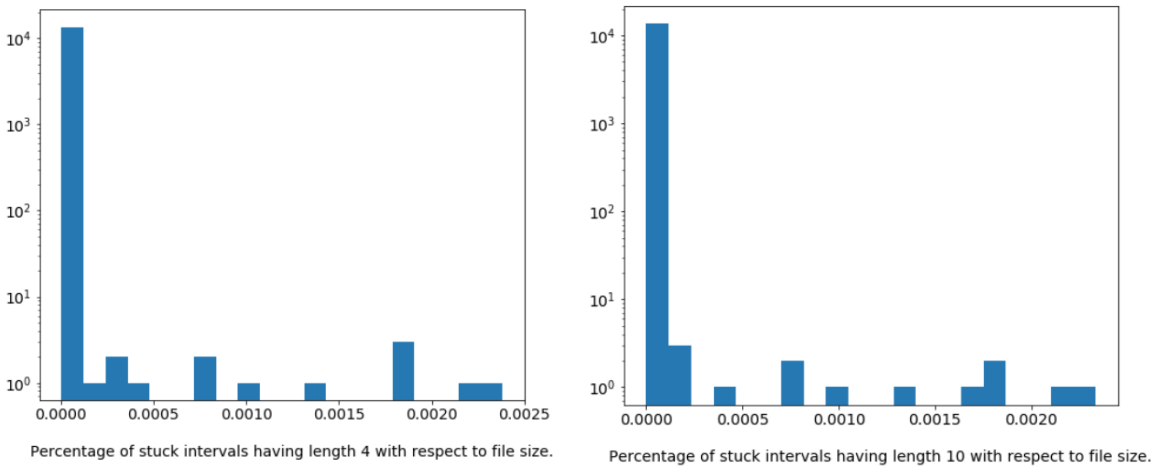


Figure 2: Percentage of stuck intervals for different time lengths

The information contained in the above plots can also be expressed as a table of percentiles:

	count	mean	std	min	25%	50%	75%	max
perc_stuck_gap1	13554.0	2.206508	0.448582	1.614220	1.696675e+00	2.575384e+00	2.595095	2.650530
perc_stuck_gap2	13554.0	0.053804	0.019371	0.028725	3.168257e-02	6.961943e-02	0.070601	0.073561
perc_stuck_gap3	13554.0	0.001301	0.000623	0.000496	5.897143e-04	1.794000e-03	0.001839	0.004201
perc_stuck_gap4	13554.0	0.000032	0.000046	0.000003	1.085714e-05	4.228571e-05	0.000047	0.002383
perc_stuck_gap5	13554.0	0.000002	0.000042	0.000000	2.857143e-07	5.714286e-07	0.000001	0.002339
perc_stuck_gap6	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002337
perc_stuck_gap7	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002337
perc_stuck_gap8	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002336
perc_stuck_gap9	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002335
perc_stuck_gap10	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002335
perc_stuck_gap11	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002334
perc_stuck_gap12	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002334
perc_stuck_gap13	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002333
perc_stuck_gap14	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002333
perc_stuck_gap15	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002332
perc_stuck_gap16	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002331
perc_stuck_gap17	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002331
perc_stuck_gap18	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002330
perc_stuck_gap19	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002330
perc_stuck_gap20	13554.0	0.000001	0.000042	0.000000	0.000000e+00	0.000000e+00	0.000000	0.002329

Table 6: Stuck interval statistics

4.2.4 Voltage statistics

Below we investigate various statistical properties of voltage and it appears the data looks okay. There are no evident calibration issues, outliers, etc.

The maximum and minimum voltage as measured at the electricity mains are expected to be c. ±340 V. We verified that the measurements in our datasets are consistent with this well-known fact. The following summary statistics were computed:

Mean of maximum voltage across all files: 345.89 V
Standard deviation of maximum voltage across all files: 6.31 V
Mean of minimum voltage across all files: -346.85 V
Standard deviation of minimum voltage across all files: 5.84 V

The histograms below summarise the minimum and maximum values of voltage across all measurements. Voltage is observed to vary between -365 V to 365 V. No 'isolated' outliers are found – rather, the distributions are found to have relatively long tails towards 'extreme' values (e.g. +365 V and -365 V). Given that there are no isolated outliers, we decided to apply a threshold of 2.5 standard deviations across this set of filters, leading to a loss of data of 0.1%.

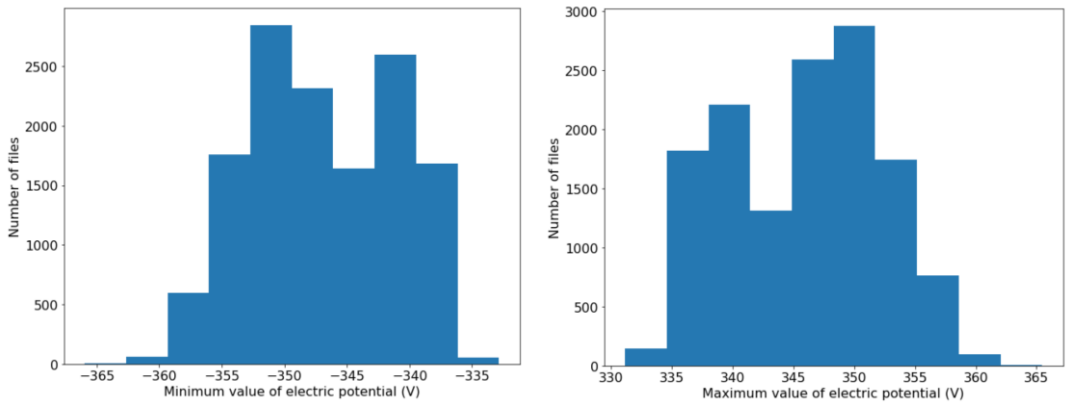


Figure 3: Histograms of min and max voltage

The root-mean square (RMS) voltage of electricity mains in the UK is quoted to be 240 V. This is a nominal value, and it is known that substantial deviations from this value can occur. This is in agreement with the following summary statistics about RMS voltage calculated from our data:

Mean of RMS voltage: 247.60 V
Standard deviation of RMS voltage: 3.84 V

The histogram below summarises the variation in RMS voltage across the whole dataset. It is worth noting that the distribution of RMS voltage has two peaks at around 243 V and around 251 V. The same behaviour is observed in the distributions of minimum and maximum voltage. This is probably due to the electricity network providing 'inconsistent service'.

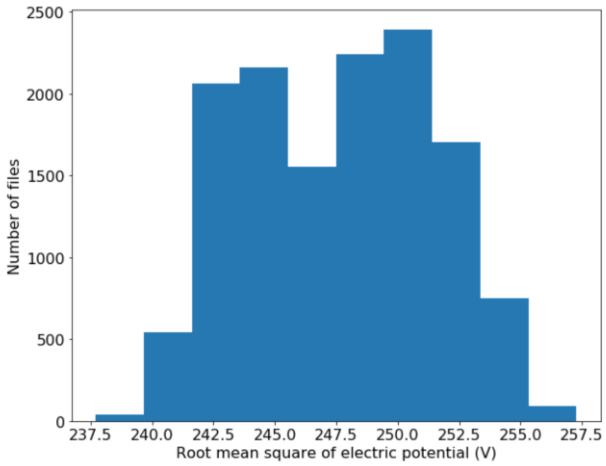


Figure 4: Histogram of RMS voltage

The RMS voltage is obviously related to the maximum value of voltage, since in the case of a perfectly sinusoidal signal the former is equal to the latter divided by the square root of two. Albeit in an approximate sense, this relationship remains true for the measurements of voltage contained in our dataset. This is demonstrated by the following histogram, representing the maximum voltage divided by the RMS voltage times the square root of two.

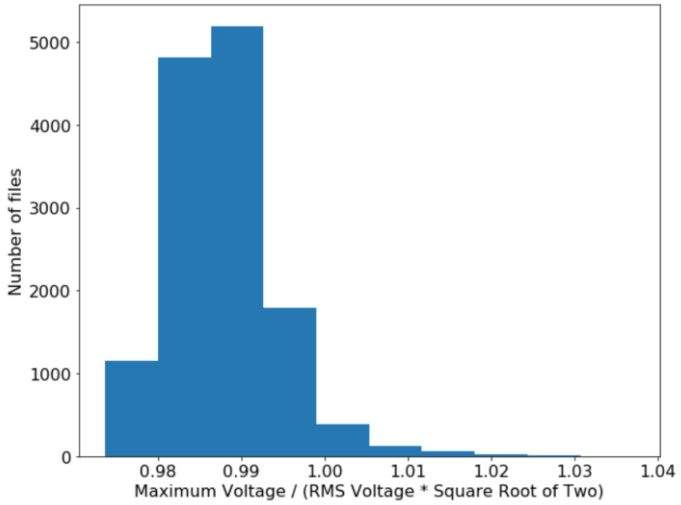


Figure 5: Histograms of *max voltage / v2 * RMS voltage*

4.2.5 Sampling rates

The reported sampling rate is always 204,918 HZ, and as such we do not remove any data for this reason.

4.2.6 Power plots

We have plotted daily load profiles and daily demand evolution over time for 4 of the 5 properties (we left out H25 as data availability was poor). These plots were a useful sense check that also highlighted that H20 had a solar panel, which was detected by the frequent negative power consumption. It is worth mentioning that some small proxies were taken when creating the plots, meaning that they are not exact, but provide a good estimate.

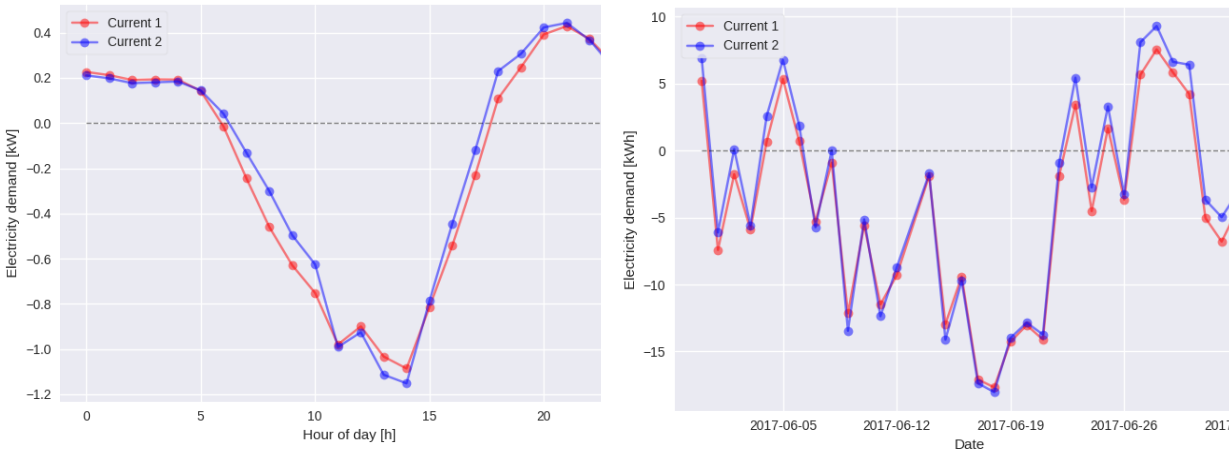


Figure 6: LHS: H20 daily load profile. RHS: H20 daily demand over time

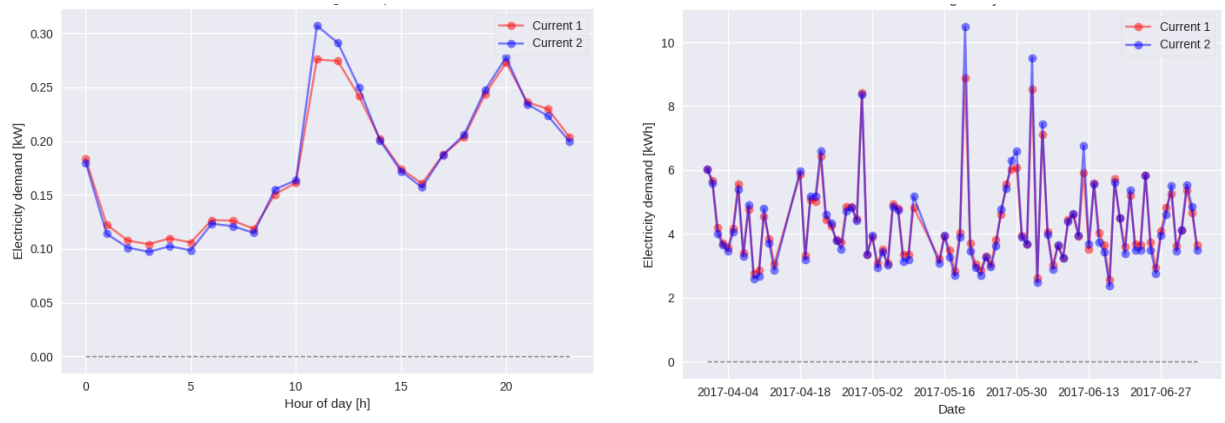


Figure 7: LHS: H45 daily load profile. RHS: H45 daily demand over time

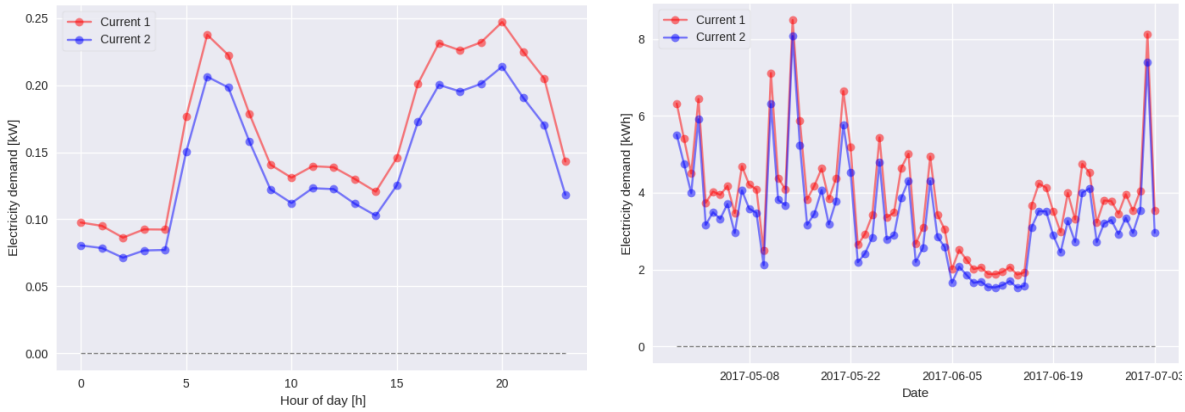


Figure 8: LHS: H71 daily load profile. RHS: H71 daily demand over time

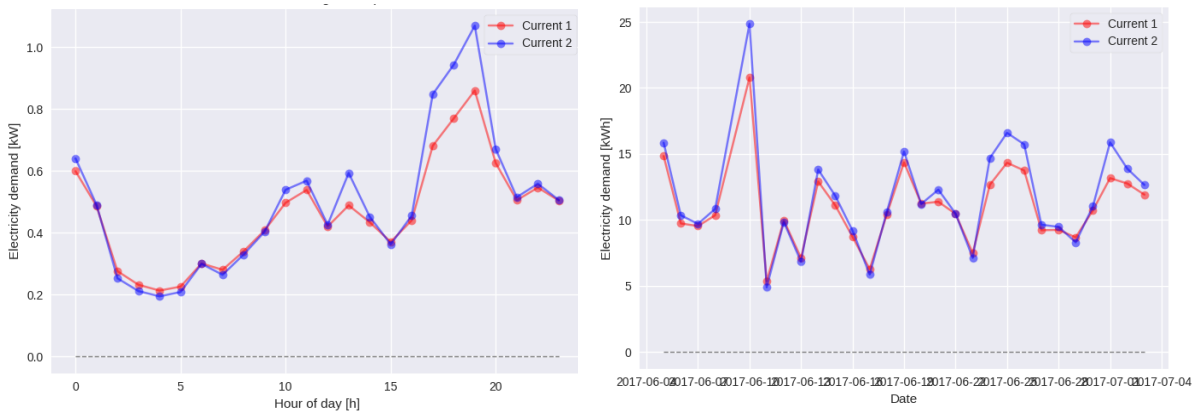


Figure 9: LHS: H73 daily load profile. RHS: H73 daily demand over time

4.2.7 Data gaps

The following plots display the number of c. 30-minute recordings gathered in distinct 6-hour intervals. It is observed that most intervals contain 12 recordings of electricity signal, as one would expect. We recall that, because the individual 30-minute files contain - with very few exceptions - the same number of datapoints, they have nearly identical durations. It appears that gaps in the electricity dataset normally last between a few days and a week. It is possible that the timing of data gaps has some consistency due to the root cause i.e. engineers do no fix outages during the weekend. We have not analysed the degree of correlation in the gaps.

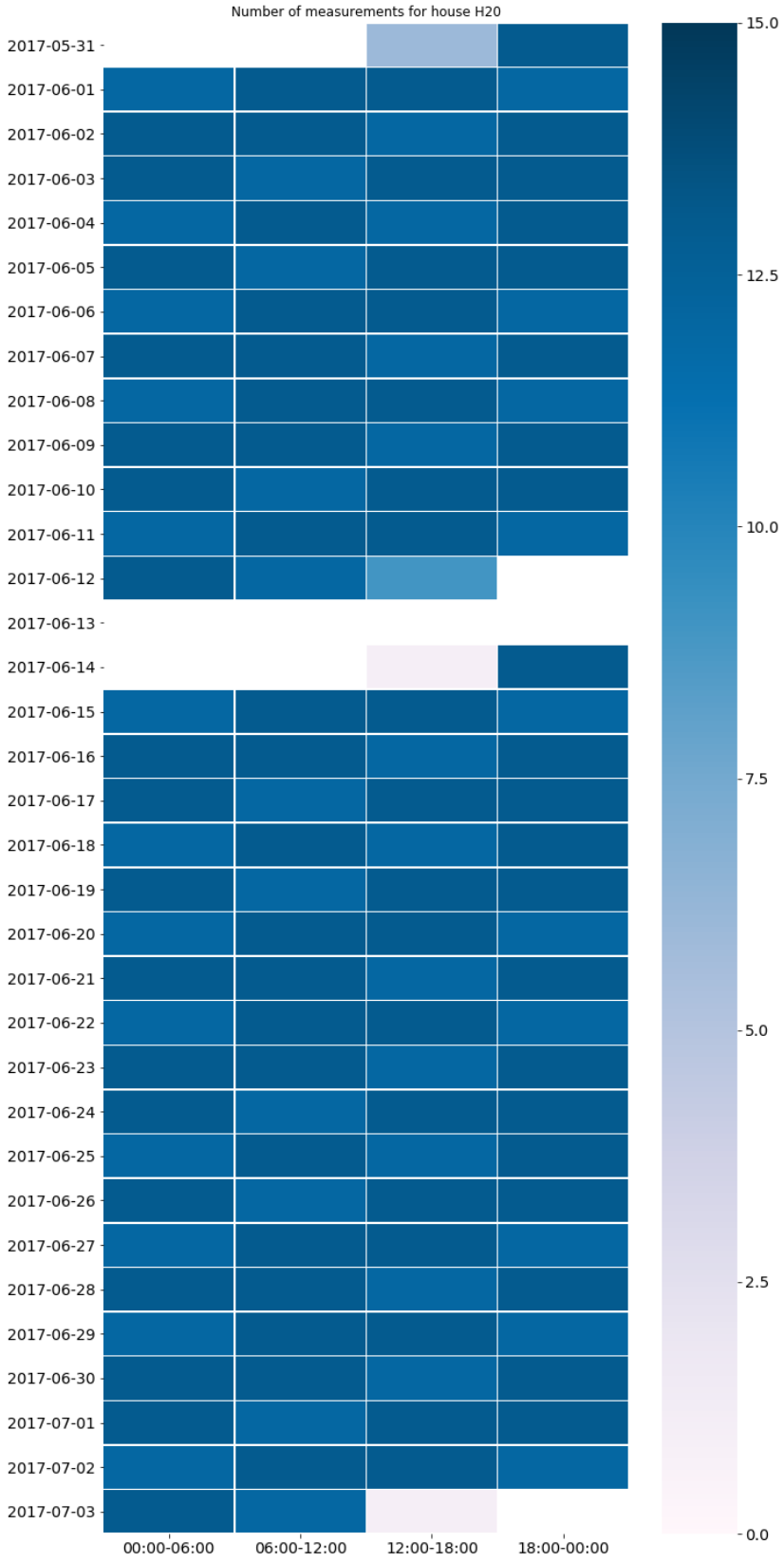


Figure 10: H20 electricity data gaps

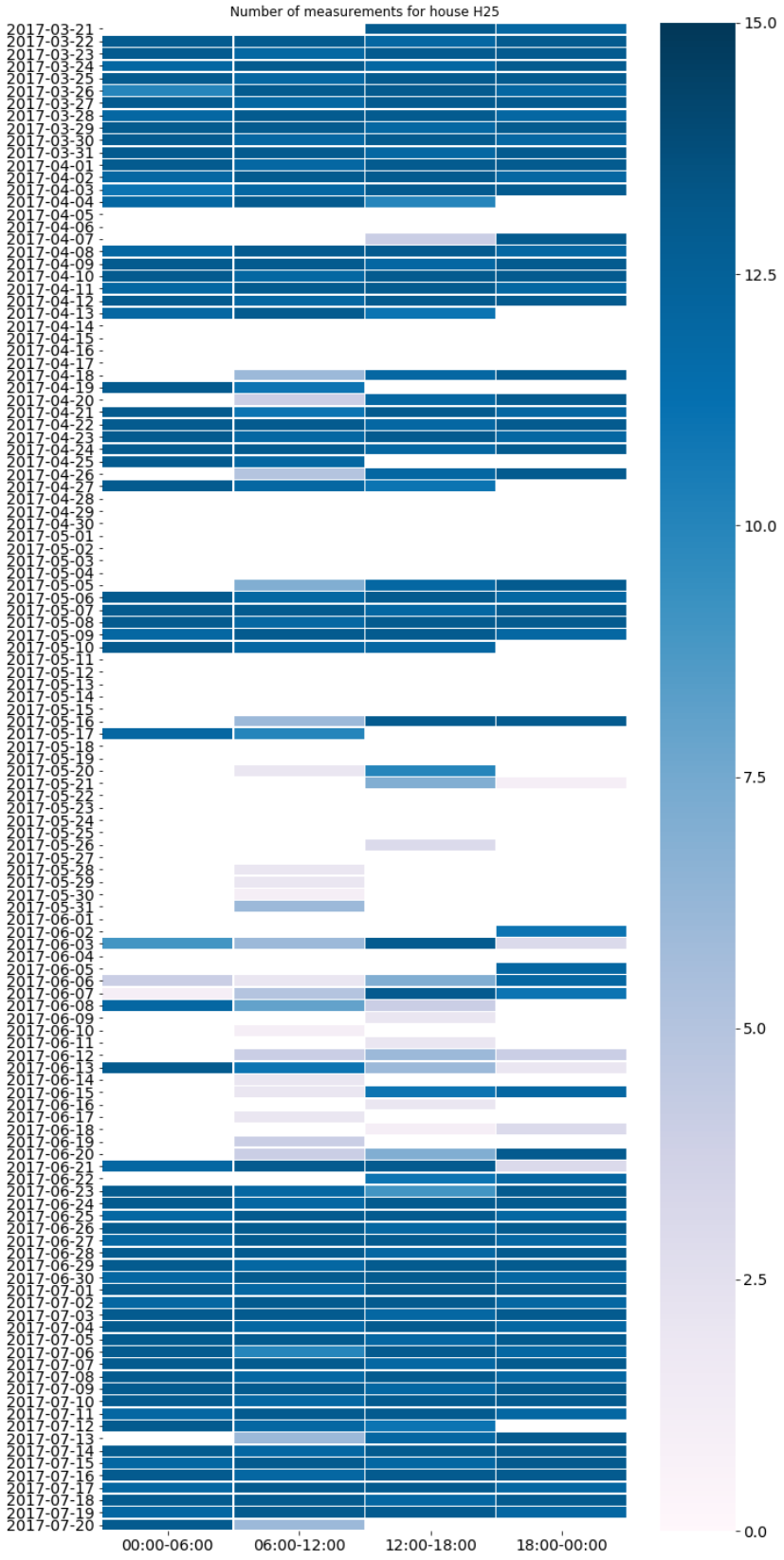


Figure 11: H25 electricity data gaps

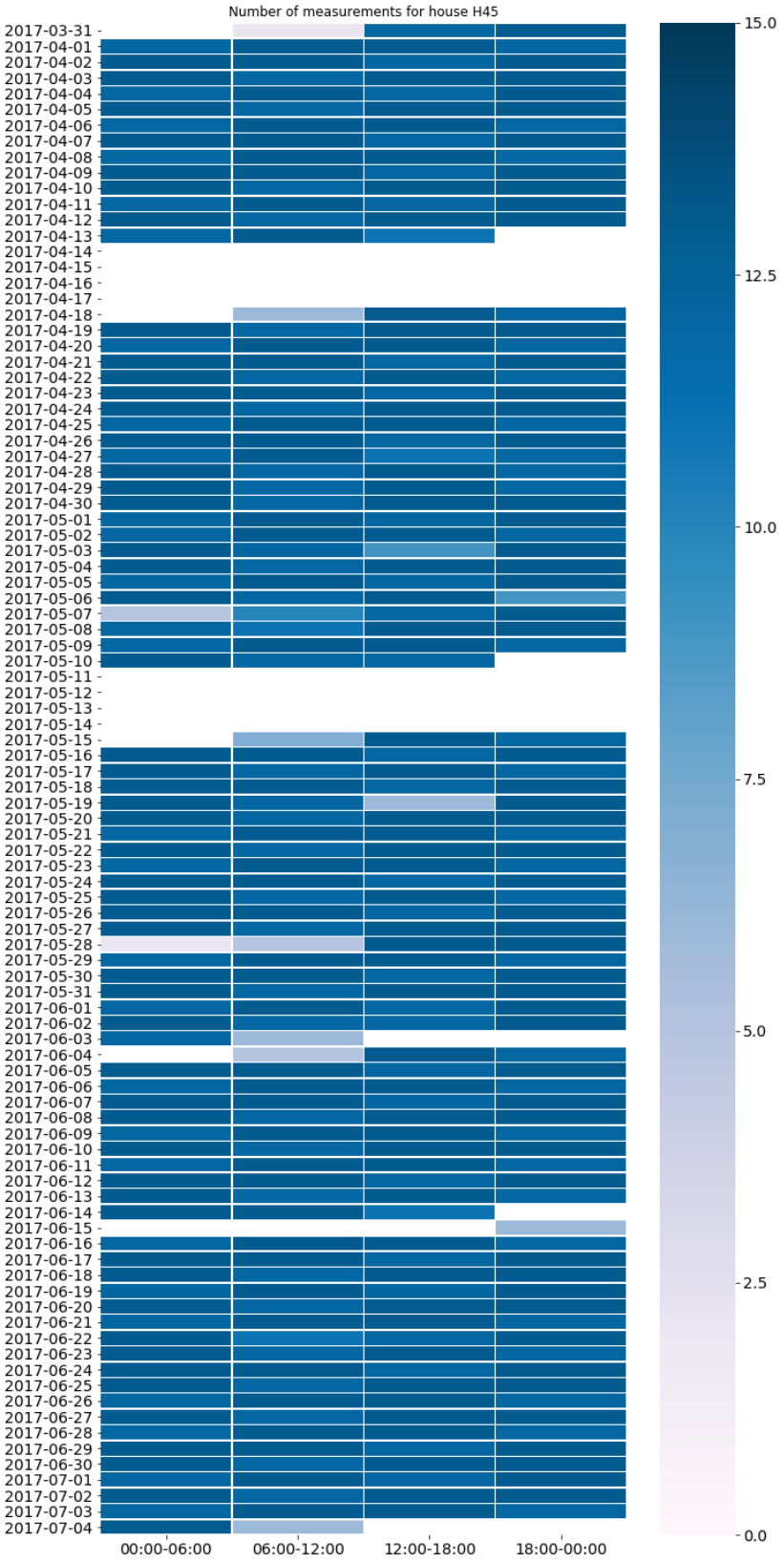


Figure 12: H45 electricity data gaps

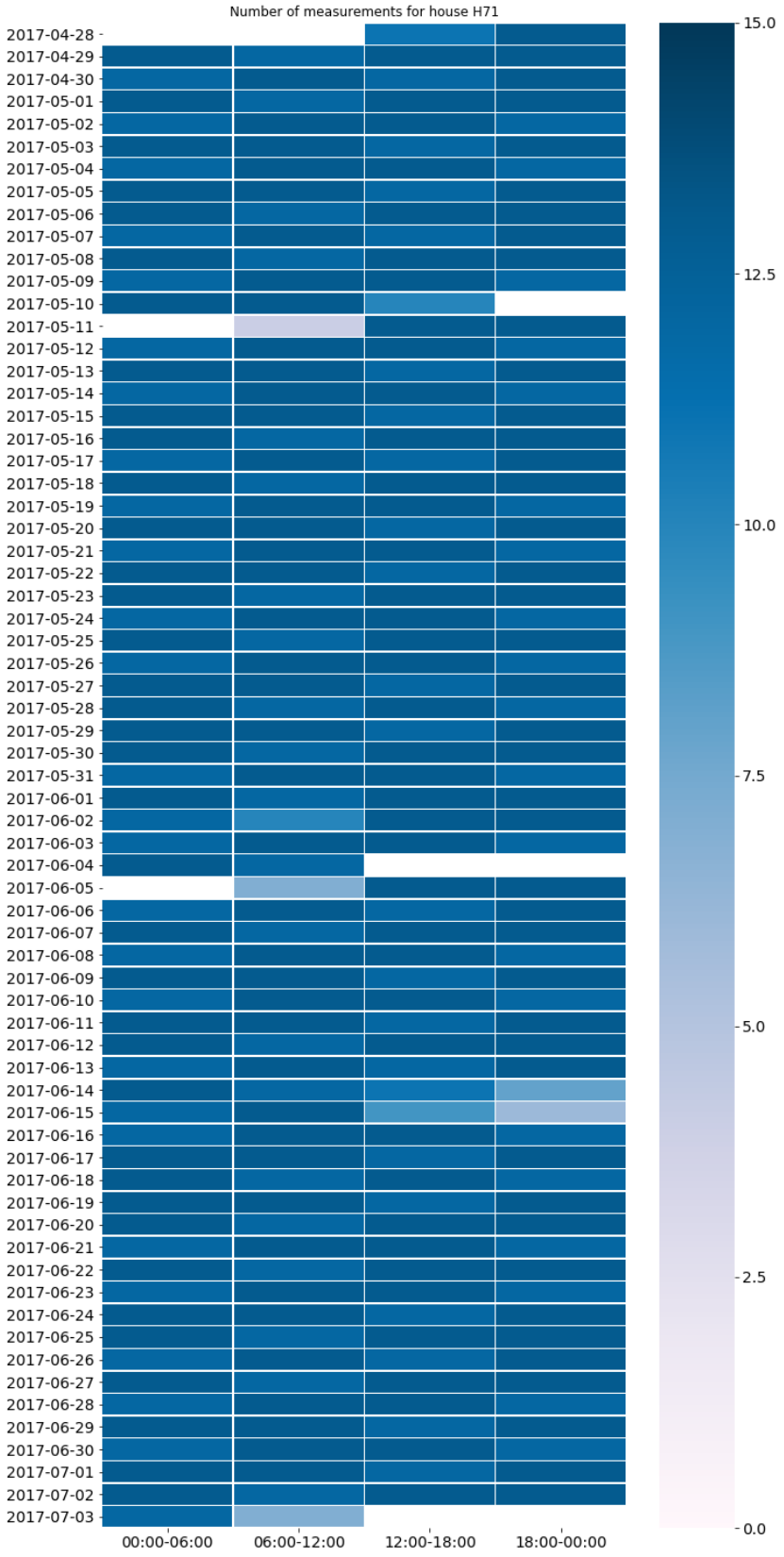


Figure 13: H71 electricity data gaps

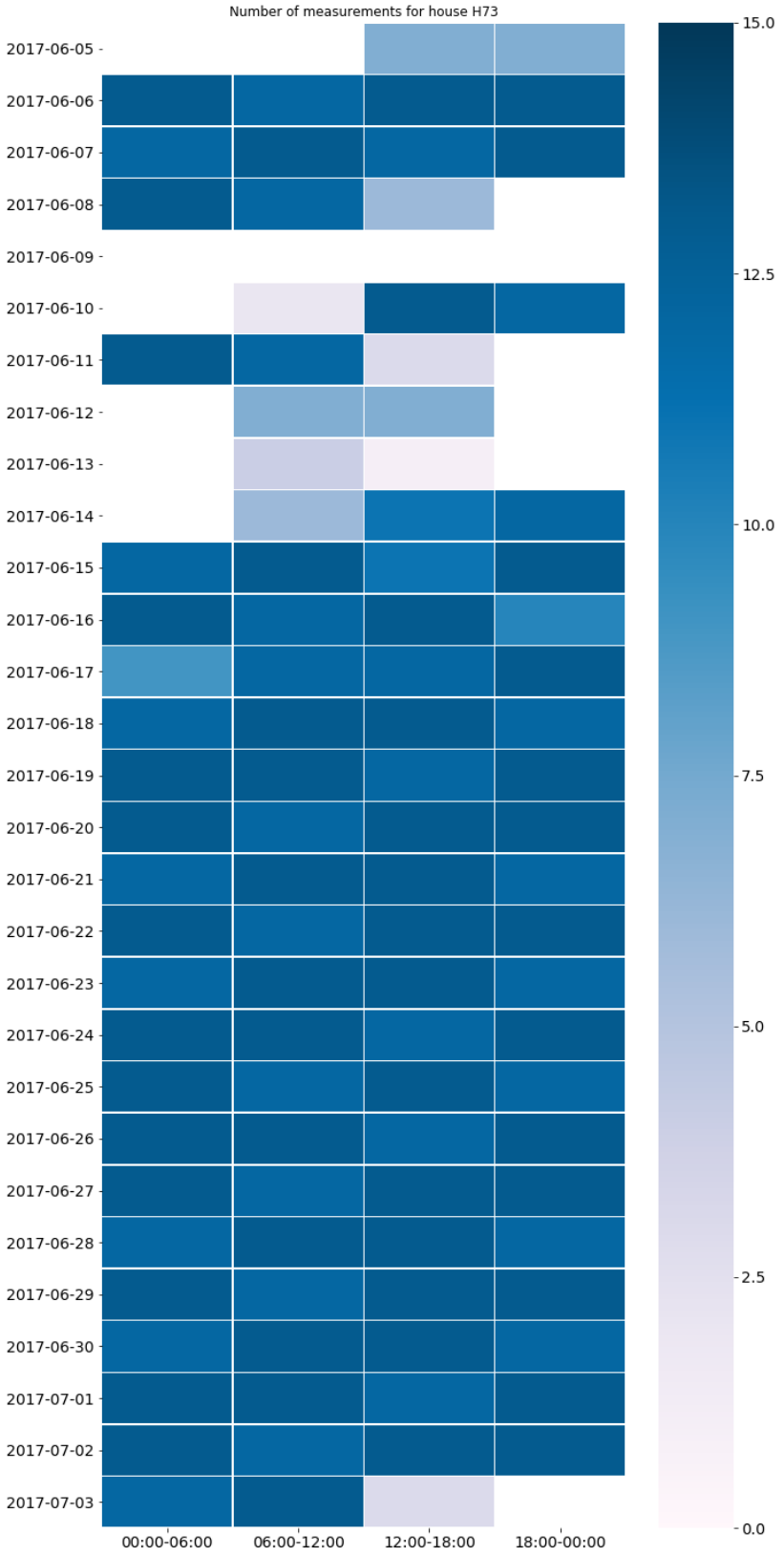


Figure 14: H73 electricity data gaps

5 Water Data Quality

Water consumption data are collected at a 10-second interval (nominal).

5.1 Overview

The overall quality of the water data is thought to be high, as observed by the following tests:

- Range of values appear reasonable
- Time drift is within the documented range
- No risk of measurement overflow
- 100% data completeness
- Average daily consumption values are reasonable

The only limitation identified is that due to the resolution around 30% of the non-zero measurements, which are of low magnitude, will have limited precision. This is discussed in more detail in Section 5.2.3. Additionally, the analysis has the limitation that it did not look into stuck values.

5.2 Detailed tests

5.2.1 Water flow statistics and histogram

In this section, water flow statistics are reported in litres per 10 seconds. An illustrative set of summary statistics is laid out in Table 7.

Statistic	Value (Litres /10 sec)
Mean	0.021
Standard deviation	0.149
Min	0.00
25th percentile	0.00
50th percentile	0.00
75th percentile	0.00
Max	2.12

Table 7: A typical set of summary statistics for water flow.

About 97.5% of the data entries are 0, which suggests that the water is on for about 2.5% of the time, corresponding to about 36 minutes per day. This amount appears reasonable.

Figure 15 shows a typical histogram of the non-zero measurements. The most frequent measurement is around 1.1 L per 10 sec, and additional distinct peaks are found around 0.1 and 1.3 L per 10 sec, respectively. The overall distribution is right-skewed and does not highlight any data quality issues.

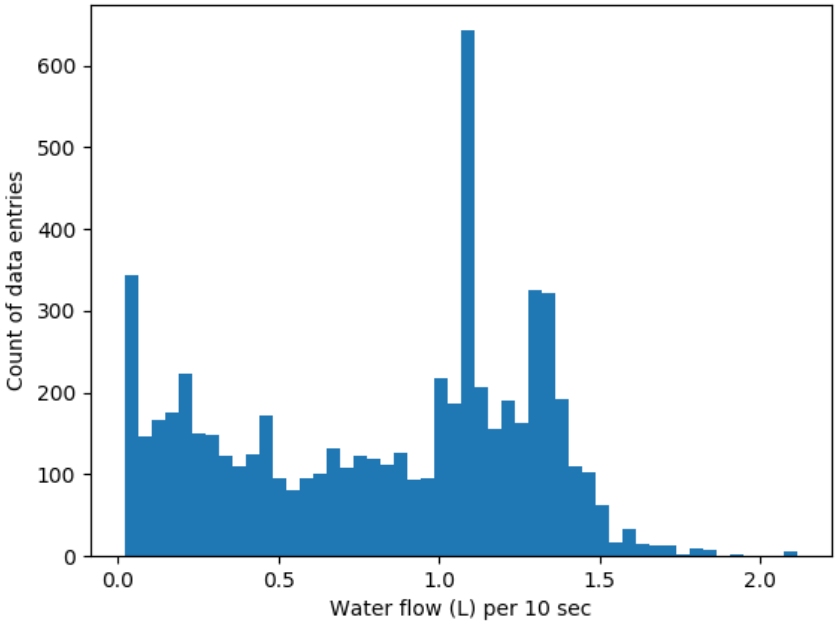


Figure 15: Histogram of non-zero water flow measurements in litres per 10 seconds.

5.2.2 Sampling rate and time drift

The water usage is sampled in 10 second intervals, and all the time drifts we have observed are within the +/- 60 seconds per month range specified in the data spec. Time drifts close to 60 seconds over 1 month are likely to pose some challenges in the analysis phase and we currently adjust for these linearly.

5.2.3 Overflow and underflow

ESC has quoted that the upper limit of the water data logger is 100 litres per minute, which should not have a risk of overflow given a typical max value is around 10 litres. The manufacturer of the data logger does not guarantee accuracy below 0.5 litres per 10 sec, which could affect the accuracy of a substantial amount of our data. In a file representing 1 month of water data in House H45, 1978 out of the 6160 non-zero data entries are observed to fall below 0.5 L per 10 sec, which means there is an accuracy risk on around 30% of the data points. No mitigation or correction methods have been provided, and therefore we leave these measurements as they are.

5.2.4 Load profile and daily demand

We examined load profiles of different properties over different time periods and found no points of concern. In Figure 16 we display the load profile for 1 month of data for H45. We can see that the occupants have different morning routines during weekdays and weekends: at weekends, consumption starts later in the day and accumulates to a higher amount. This agrees with the life style of a common family.

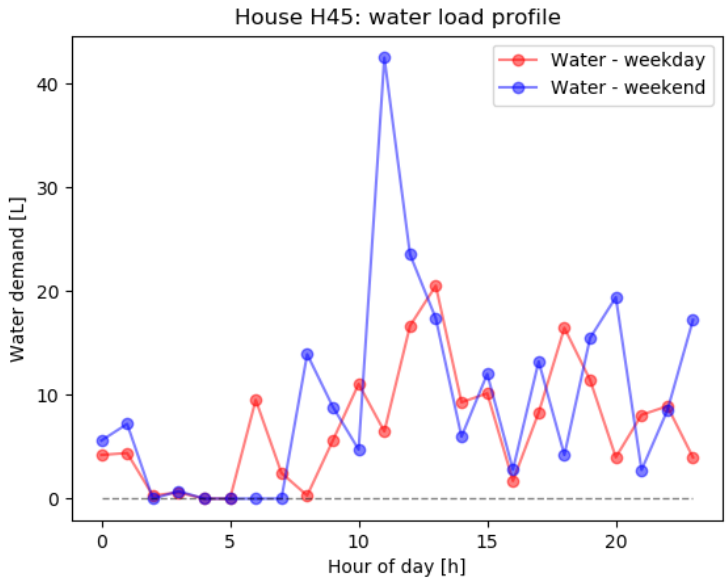


Figure 16: Load profile averaged from 1 month of water data in House H45

Figure 17 plots the daily water consumption in House H45 by day of week over the month, where the usage shows a strong weekly pattern. For example, Mondays, Tuesdays and Wednesdays of weeks 14-16 show similar consumption amounts. The daily consumption varies around 200 L, with a few extremely low values around 50 L/day on Fridays and an extremely high value of 400 L on one Monday. Both the weekly pattern and the consumption figures are considered within expectation.

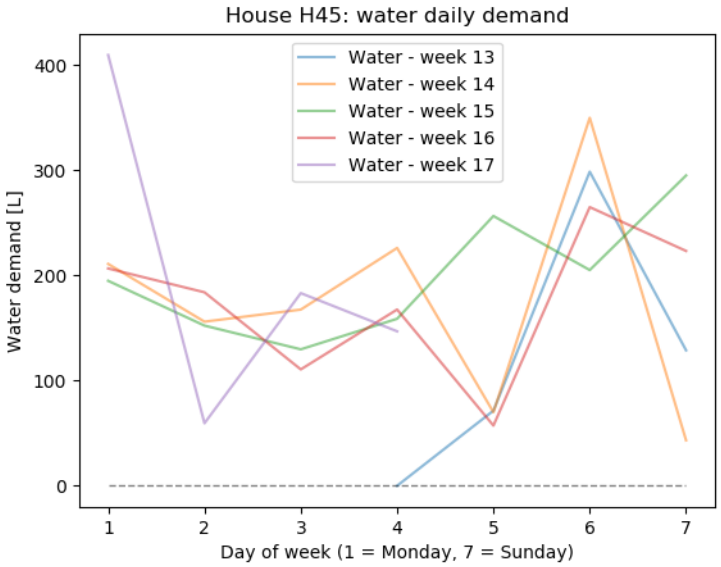


Figure 17: Daily water demand (consumption) by day of week in House H45

A more granular view of the above plots is below. The highest value of 400 L / day identified in Figure 17 is represented by the darkest square in the lower right quadrant of Figure 18. It shows that more than half of the water consumption on that Monday was consumed between 6-7pm, while the rest of the day just followed the pattern of other Mondays. Other relatively dark squares are consistent with the peak in the load profile curve for weekends as shown in Figure 16.

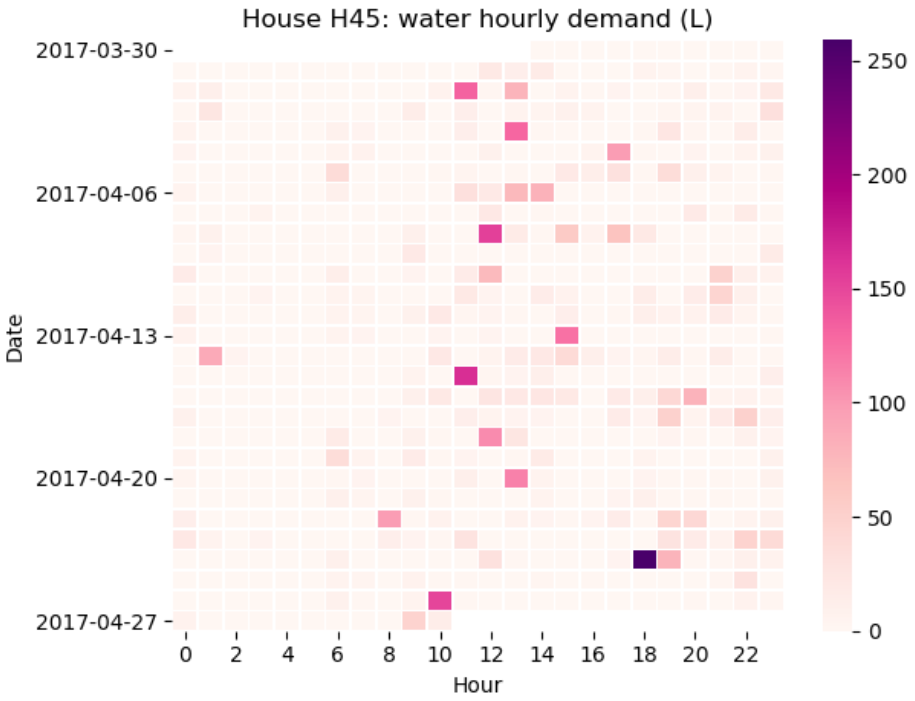


Figure 18: Water demand by day and hour presented by heatmap

5.2.5 Data gaps

The water data provided have very good quality in terms of completeness; no data gaps are observed.

6 Gas Data Quality

6.1 Overview

The gas data was extracted from the HEMS MongoDB and covers April - May 2017 for 3 properties. The gas data presented a few challenges:

1. The gas flow is sampled on a much lower frequency than the other datasets (the nominal interval between samples is 15 minutes), leading to a low resolution and syncing challenges.
2. The connection between the gas meter and the HEMS Hub drops on occasion. The 4 connection types are listed in the table below. Following the advice of ESC, our analysis is based solely on datapoints marked “Normal”.
3. The unstructured nature of the MongoDB Schema presented a few challenges, making it harder to extract data.

Connection type	Description
Connecting	Initially, sensors are registered with the HUB, and the HUB starts issuing a connecting status, until the sensor is connected to the HUB.
Warning	Either the sensor failed to communicate with the HUB for 15 minutes, or if the sensor hasn't communicated for two consecutive periods (period is dependent on the sensor resolution, some sensors report every 1 minute, every 5 minutes, etc.), then the HUB issues an absent status
Absent/Failed	If the sensor failed to communicate with the HUB for 24 hours consecutively
Normal	HUB is connected with the sensor, and the sensor is reporting some readings.

Table 8: Description of four connection types

An important detail about the HEMS dataset is that many samples are duplicated. Conversations with ETI confirmed that these duplicates are generated when the HEMS hub attempts to upload data to the cloud-based server. We suspect that the detailed mechanism is as follows. The hub keeps sending a specific measurement to the cloud server until it receives confirmation that the upload was successful. At times, this confirmation from the server fails to arrive at the hub, forcing it to attempt a new upload, in spite of the data-point having actually being stored on cloud. This results in measurements being recorded multiple times. Following the advice of ETI, we started our analysis by removing such duplicate measurements. It is worth mentioning that the analysis has the limitation that it did not look into stuck values.

6.2 Detailed tests

6.2.1 Summary statistics and histogram of gas usage

Summary statistics of gas flow (usage) across 1.5 months in House H45:

Statistic	Value
Mean	0.00782
Standard deviation	0.0365
Min	0
25th percentile	0
50th percentile	0
75th percentile	0
Max	1.48

Table 9: Summary statistics of gas flow

We observe that gas is only used 24% of the time, which is a reflection that few appliances utilise gas. The plot on the left-hand side of Figure 19 provides the histogram of non-zero gas flow measurements. We note that, as the value of gas flow increases, the corresponding number of events decreases quickly. This signifies that events with large gas consumption are rare. The plot on the right-hand side offers a zoomed-in view on small values of gas flow.

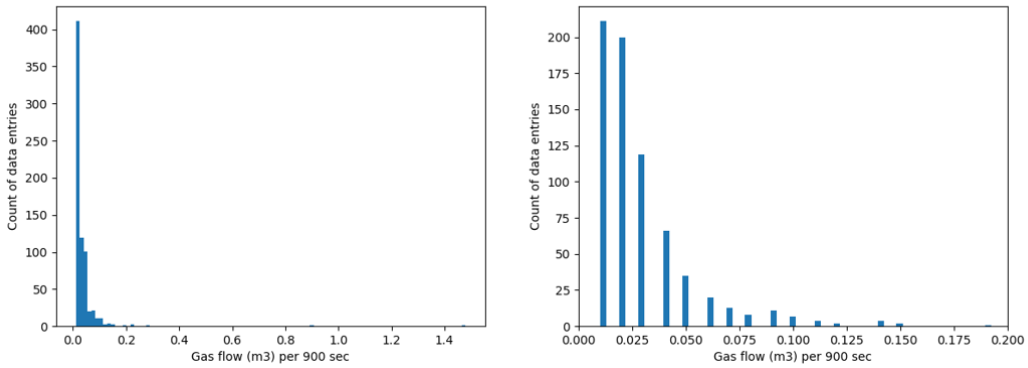


Figure 19: (Left) Histogram of non-zero gas flow in m^3 per 900 seconds for house H17. (Right) oomed-in histogram.

6.2.2 Data gaps and duplicates

The gas dataset comes with rather frequent gaps and missing samples. These are due to the gas meter failing to connect to the HEMS hub, as well as to other causes that cannot be identified. Typically, around 30% of the data entries are missing.

Figure 20, **Figure 21** & Figure 22 indicate how many measurements were collected in a given hour for the houses H25, H45 and H71, respectively. White squares correspond to data gaps with duration of at least one hour.

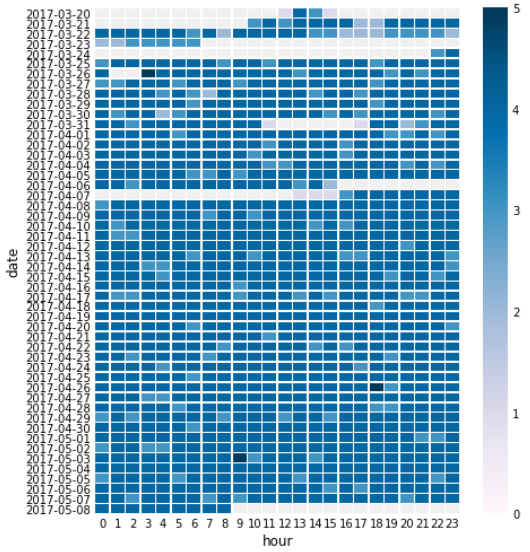


Figure 20: Number of gas measurements per hour in the house H25.

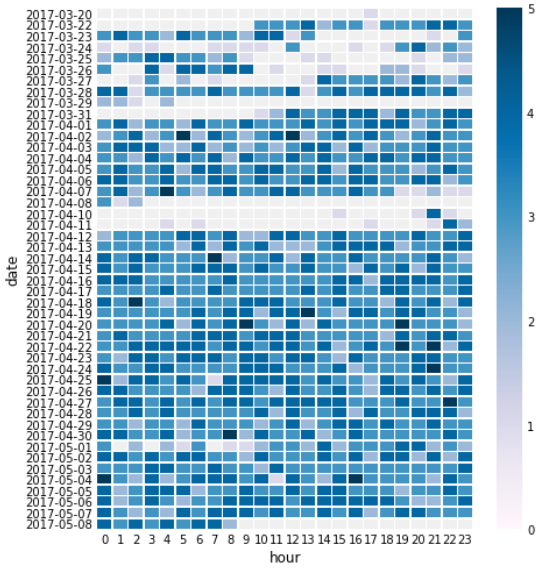


Figure 21: Number of gas measurements per hour in the house H45.

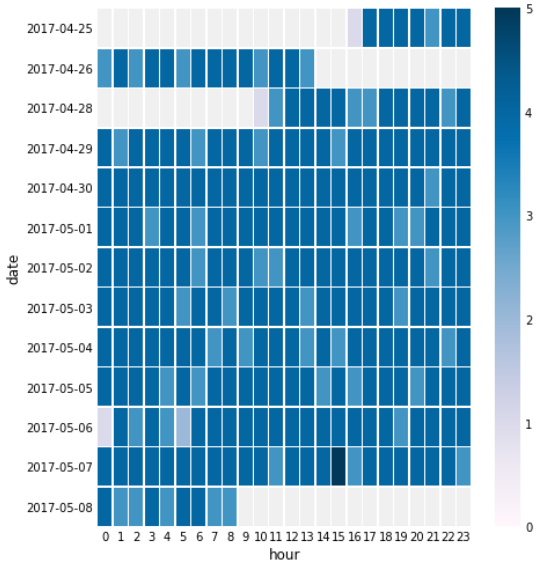


Figure 22: Number of gas measurements per hour in the house H71.

6.2.3 Daily demand and load profile

Figure 23, Figure 24 and Figure 25 represent the consumption of gas on a given hour for the houses H25, H45 and H71. Red blocks correspond to hours where the gas usage is high, light yellow blocks correspond to hours where the gas usage is low, and white blocks correspond to data gaps.

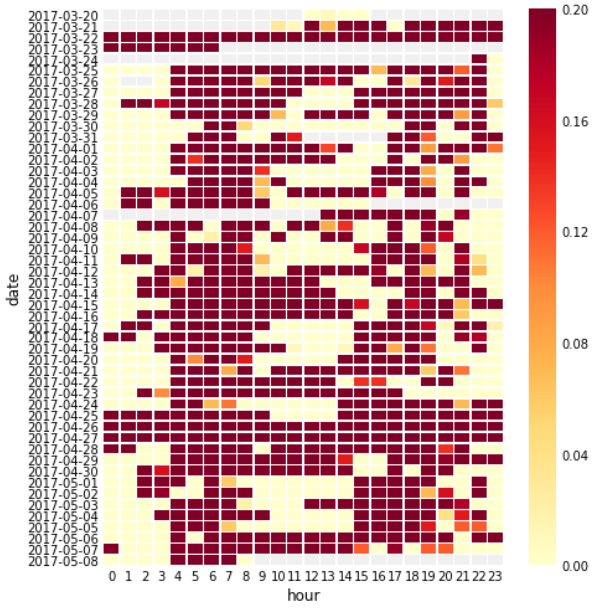


Figure 23: Consumption of gas per hour in the house H25.

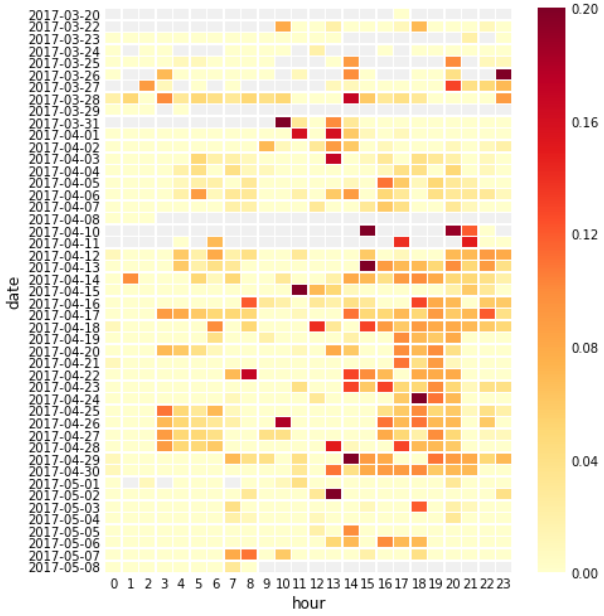


Figure 24: Consumption of gas per hour in the house H45.

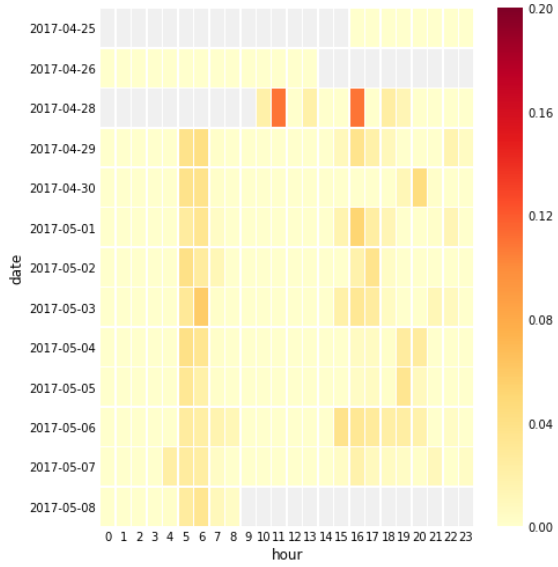


Figure 25: Consumption of gas per hour in the house H71.

As one would expect, gas consumption displays peaks in the morning and in the evening, when heating is likely to be on. Moreover, different consumption patterns are observed for the weekdays and the weekend, see Figure 26.

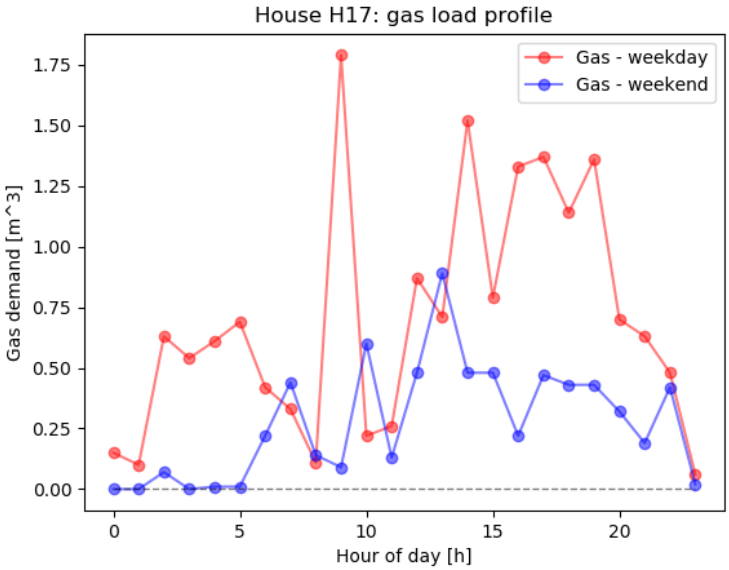


Figure 26: Average hour consumption for each hour of the day.

7 Temperature and Humidity Data Quality

7.1 Overview

The HEMS data that we have focused on, other than gas, is temperature and humidity, as we believe these are likely to be key to our analysis. Our tests, suggest that there are no major issues with the current data. The hourly profiles analysed suggest that humidity is likely to be a better indicator of resident needs, and help identify meals and bathing. Temperature and humidity are measured in all rooms of each property at intervals of 1 and 5 minutes, respectively. It is worth mentioning that the analysis has the limitation that it did not look into stuck values.

7.2 Detailed tests

7.2.1 Data gaps

Figure 27, Figure 28 & Figure 29 display the number of samples collected over a given hour in properties H25, H45, and H71, respectively. More specifically, the plots refer to the bathroom in each house that is regularly used. In the dataset, this room is labelled “loo” for H25, “bathroom” for H45, and “en-suite” for H71. The graphs on the left-hand side of Figure 27, Figure 28 & Figure 29 correspond to temperature data, and the graphs on the right-hand side correspond to humidity data.

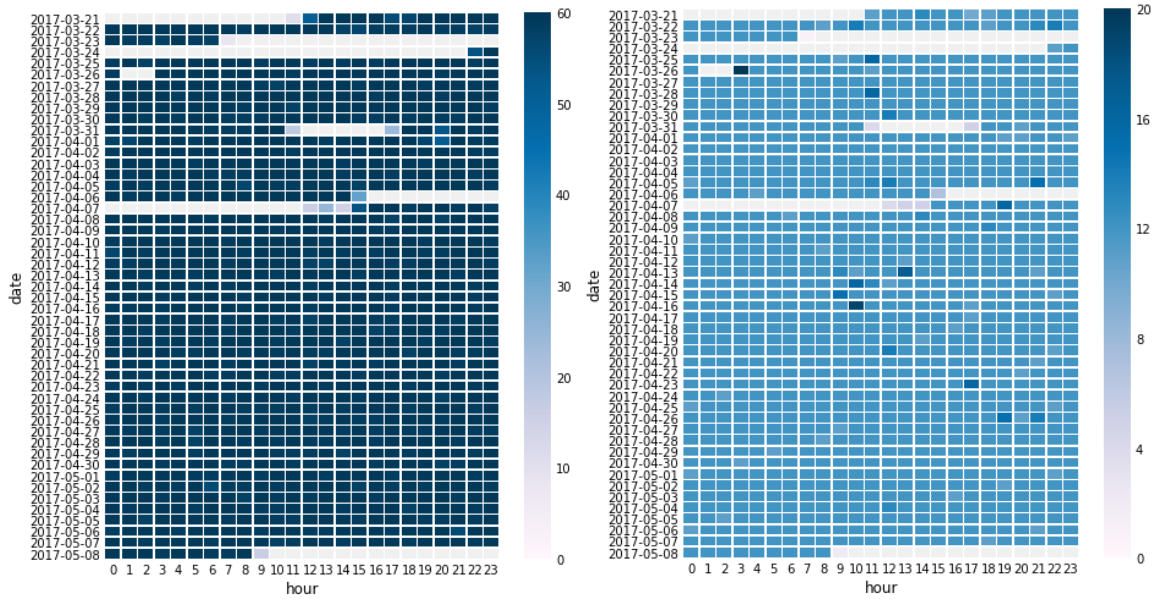


Figure 27: Number of temperature (Left) and humidity (Right) measurements per hour in the house H25 (Room “Loo”).

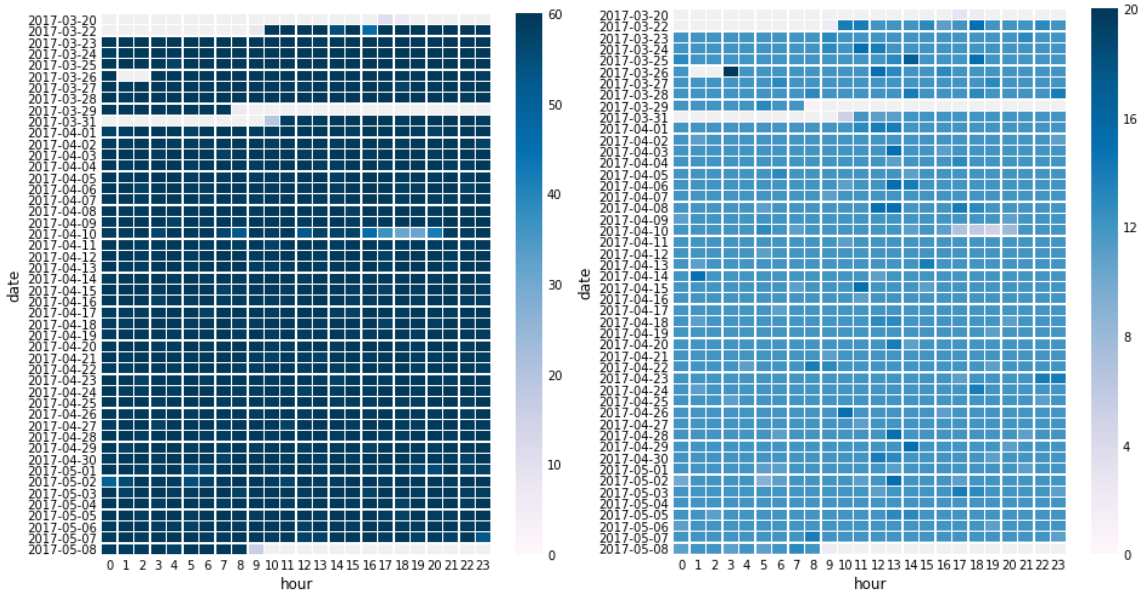


Figure 28: Figure 17b: Number of temperature (Left) and humidity (Right) measurements per hour in the house H45 (Room "Bathroom").

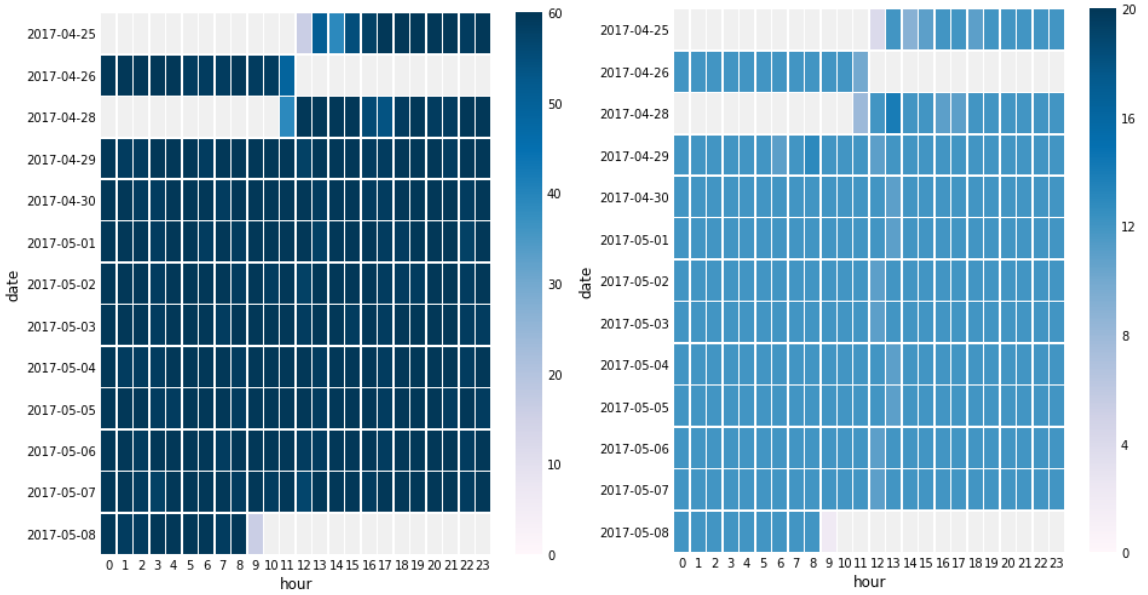


Figure 29: Number of temperature (Left) and humidity (Right) measurements per hour in the house H71 (Room "En suite").

7.2.2 Daily demand

Figure 30 & Figure 31 & Figure 32 depict temperature and humidity measurements over a given hour in properties H25, H45, and H71, respectively. As in the previous section, the graphs on the left-hand side represent temperature measurements, and the graphs on the right-hand side represent humidity measurements.

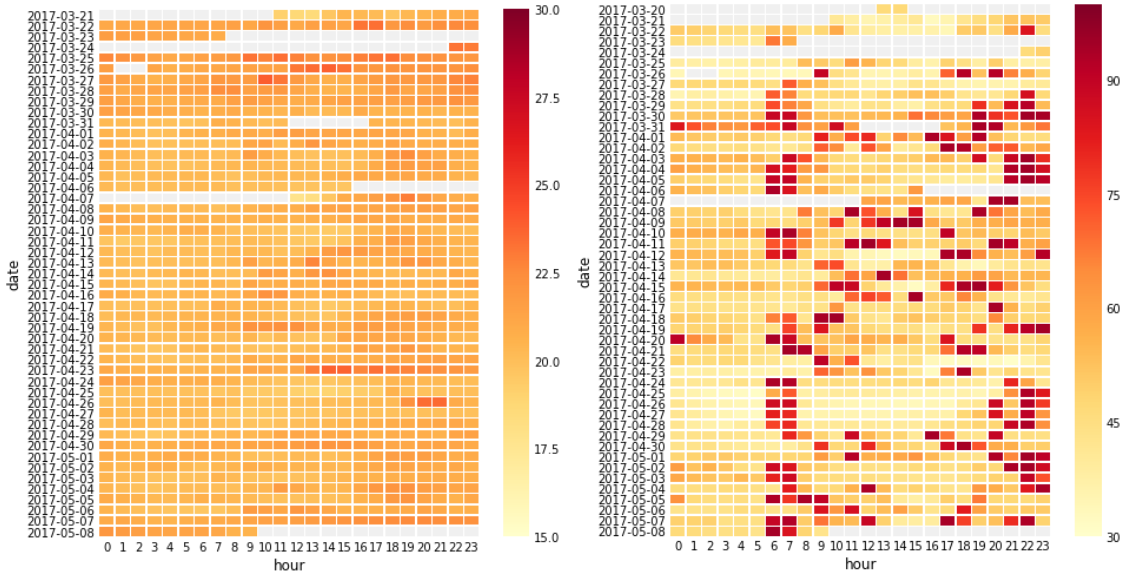


Figure 30: Average temperature in degrees Celsius (Left) and humidity in percentage (Right) for house H25 (Room: "Loo").

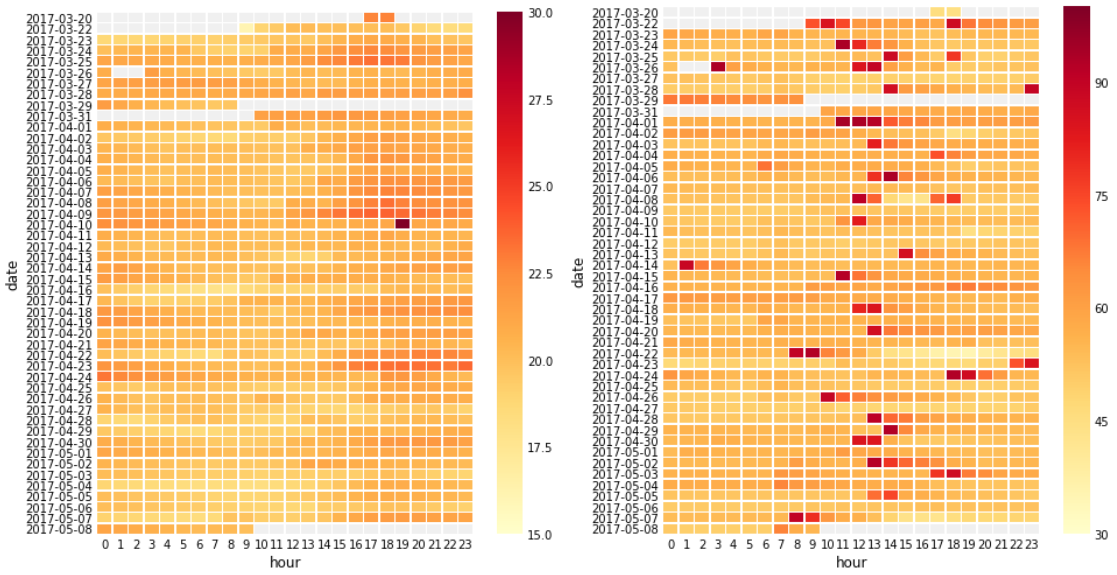


Figure 31: Average temperature in degrees Celsius (Left) and humidity in percentage (Right) for house H45 (Room: "Bathroom").

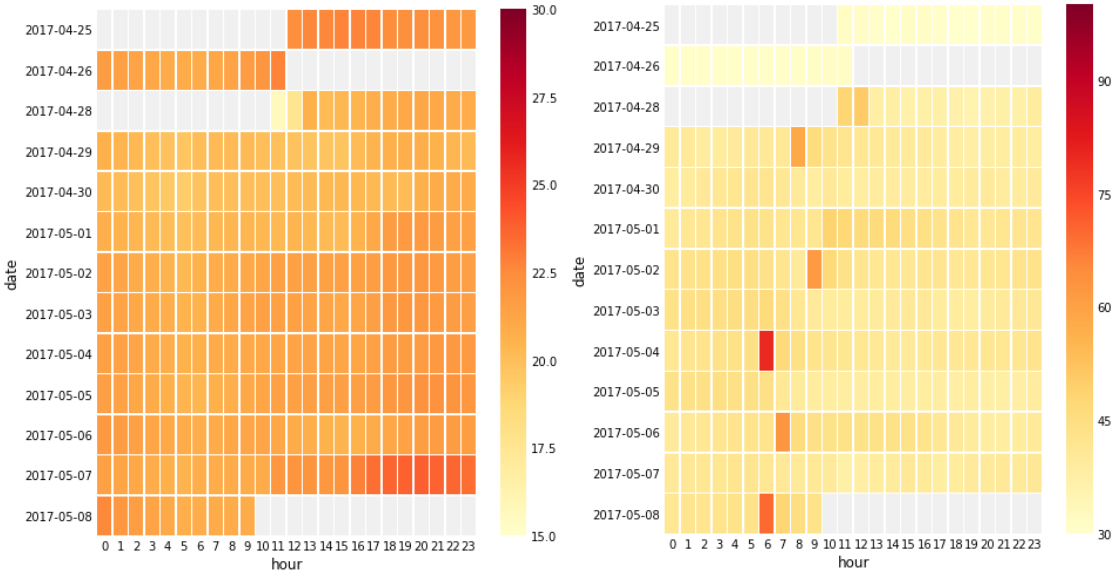


Figure 32: Average temperature in degrees Celsius (Left) and humidity in percentage (Right) for house H71 (Room: "En suite").

In Figure 30 & Figure 31 & Figure 32, changes in humidity are very evident, whereas changes in temperature are gradual and less noticeable. One deduces that humidity measurements are in all likelihood better predictors of human actions than temperature measurements. For example, recurrent high values of humidity in the bathroom can be easily interpreted as showering. Figure 30 indicates that the inhabitants of H25 typically take two showers a day, whereas Figure 31 indicates that the inhabitants of H45 typically take one shower a day.

8 Conclusions and Recommendations

The work to date has focused on generating an inventory of the data, as well as running various statistical tests and sense checks. The results of this suggest that due to data availability our initial efforts are best focused on April data for property H45, which is a property with minimal apparent quality issues and has the highest amount of data.

	H20	***H25***	***H45***	***H71***	***H73***
No. of hard drives	1	3	3	2	1
Electricity data timespan	31 May – 3 Jul	21 Mar – 20 Jul	31 Mar – 4 Jul	28 Apr – 3 Jul	5 Jun – 3 Jul
Electricity data quality	Solar panels present	Frequent, repeated long gaps	~15 days missing; well-defined gaps	~2 days missing; well-defined gaps	~5 days missing; well-defined gaps
Water	31 May – 3 Jul	21 Mar – 20 Jul (2 water meters)	30 Mar – 4 Jul. Good data quality	28 Apr – 3 Jul. Good data quality	5 Jun – 3 Jul. Good data quality
HEMS database 1	NA	20 Mar – 8 May	20 Mar – 8 May	25 Apr – 8 May	NA
HEMS database 2	Unaudited	Unaudited	Unaudited	Unaudited	Unaudited
Home survey & floor plans	Available	Available	Available	Available	Available

Table 10: Overview of data received for the five properties.

The key findings to date are:

1. Electricity data appears to be of a high quality (will remove c. 3%) with the exception of some significant recording gaps;
2. Water data quality is very high with the exception of two issues: a) time drift of close to 60 seconds over 1 month, and b) small consumption values are below the manufacturers recommended minimum for high accuracy;
3. HEMS data is in an inconvenient format for processing, adding additional work.
4. Gas data quality was poor and poses various challenges, which include syncing due to low frequency and a high amount of missing data to an unstable connection;
5. Temperature and humidity did not have any apparent major issues and humidity is likely to be a key data feature;
6. Property H20 is harder to analyse as it has a solar panel;

Overall, data quality issues are not a blocker, but do complicate the analysis and make some of the data unusable. As we scale the analysis, data quality will be a key consideration.